Repräsentanz und Data Mining - Konzepte und Methoden der digitalen bodenkundlichen Kartierung

Dissertation
zur Erlangung des Grades eines Doktors der Naturwissenschaften

der Geowissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von

Karsten Schmidt

aus Burg

2009

Tag der mündlichen Prüfung: 23.07.2009

Dekan: Prof. Dr. Peter Grathwohl

1. Berichterstatter: Prof. Dr. Thomas Scholten

2. Berichterstatter: Prof. Dr. Volker Hochschild

Inhalt

1 Einleitung

Vor dem Hintergrund des globalen Klimawandels, einer stetig wachsenden Weltbevölkerung und einer begrenzten Ressourcenverfügbarkeit nimmt der Druck auf die Ressource Boden als Grundlage für die Nahrungsmittelproduktion zu (Corwin and Lesch, 2005). Um entsprechende Handlungsempfehlungen für eine nachhaltige Nutzung von Böden erstellen zu können, die den vielseitigen Nutzern - von Seiten der Land- und Forstwirtschaft, des Hochwasserschutzes, des Landschaftsschutzes sowie der räumlichen Planung - gerecht zu werden, bedarf es möglichst detaillierter flächendeckender, digitaler Bodeninformationen.

Gegenläufig zum weltweit steigenden Bedarf an hochaufgelösten Bodenkarten steht der allgemeine Abbau an Finanz- und Personalmitteln in den zuständigen Geologischen Diensten (McBratney et al., 2002; Behrens und Scholten, 2006a). Daher stehen für die detaillierte bodenkundliche Landesaufnahme kaum noch freie Ressourcen zur Verfügung. In diesem Spannungsfeld ist in den letzten Jahren eine starke Nachfrage nach mathematischen und statistischen Verfahren zur Regionalisierung von Bodendaten erwachsen. International hat sich in diesem Zusammenhang das Forschungsfeld „Digital Soil Mapping“ etabliert, das als Schnittstelle zwischen Grundlagenforschung und Anwendung gesehen werden kann (Lagacherie et al., 2006). Ziel des „Digital Soil Mapping“ ist es, mit Hilfe von raumstatistischen Verfahren valide Inter- und Extrapolationen zu ermöglichen. Dabei sollen - idealerweise mit relativ kleinen Stichprobenumfängen - relativ große Flächen hochauflösend mit bodenkundlichen Informationen versehen werden. Dies führt zu einem gestiegenen Bedarf an effizienten Methoden und Algorithmen zur Datenanalyse und Extrapolation. Durch die rasante Entwicklung der Rechentechnik in den letzten 10 Jahren ist es möglich geworden, große Datenbestände auf Standardhardware zu analysieren und somit hochauflösende Prognosen über die Feldskala hinaus zu ermöglichen. Für die Boden-Landschaftsmodellierung sind hier insbesondere Methoden des Data Mining für die Anwendung in Geographischen Informationssystemen im Allgemeinen bzw. in der Bodenkunde im Speziellen (McBratney et al., 2003; Scull et al., 2003; Behrens et al., 2005; Bishop et al., 2006) prädestiniert, da sie für die Identifizierung komplexer räumlicher Zusammenhänge (Jenny, 1941; McBratney et al., 2003) und die Ableitung von Regel- oder Formelwerken in unscharfen Datenmengen konzipiert sind.

Die Anwendung von Regressionsverfahren und Verfahren der überwachten Klassifikation in der Boden-Landschaftsmodellierung erfordert eine Lerngrundlage (Stichprobe), die den Merkmalsraum einer Landschaft in all seinen Ausprägungen detailliert widerspiegelt. Die Größe der Stichprobe, die für eine systematische und vollständige Abdeckung der Landschaft benötigt wird, ist dabei von der Diversität bzw. Heterogenität der jeweiligen Landschaft abhängig. Generell gilt, dass mit steigender Gebietsgröße und höherer Heterogenität mehr Proben für eine hochauflösende Modellierung benötigt werden. Um in heterogenen Landschaften valide

Bodenprognosemodelle erstellen zu können, empfiehlt es sich die Landschaft zu segmentieren und die einzelnen homogeneren Bereiche separat zu modellieren (McBratney et al., 1991). Zur systematischen Segmentierung ist es im Rahmen der Bodenlandschaftsmodellierung daher notwendig, Verfahren zur Landschaftsstrukturanalyse in den Prozess der Modellierung zu integrieren. Dies ist insbesondere für Studien mit räumlichen Auflösungen ≤ 20 m und Landschaften > 1000 km² der Fall. Für solche Gebietsgrößen müssen nicht nur repräsentative Beprobungsschemata und -verfahren entwickelt werden, sondern es sollte auch der Einsatz von geophysikalischen Naherkundungstechniken in Betracht gezogen werden (McBratney et al. 2000, 2003; Behrens und Scholten, 2006a; Viscarra-Rossel et al., 2007; Gerber et al., 2008), um den Beprobungsaufwand zu minimieren (McBratney et al., 2006).

Ein weiterer wichtiger Punkt im Rahmen von Data Mining-basierten Studien zur Bodenprognose im Landschaftsmaßstab ist die enorme Menge anfallender Daten sowie die Qualität der zur Verfügung stehenden Basisdaten. Letztere sind meist nicht als direkte Grundlage für räumliche Modellierungen erstellt worden und unterliegen als so genannte „Altdaten“ verschiedenen Beschränkungen in Bezug auf ihre Qualität und Interpretierbarkeit. Im Rahmen der digitalen Bodenkartierung muss daher in vielen Situationen mit Datensätzen umgegangen werden, die sich stark in Bezug auf Herkunft, Alter, Maßstab, Auflösung, und Kartierungsstrategie unterscheiden, was zu verschiedensten Fehlerquellen führen kann (Robinson et al., 1984; Lagacherie und Holmes, 1997; Heuvelink, 1998; Bishop et al., 2006; Behrens et al., 2008). Werden beispielsweise mittel- oder kleinmaßstäbige Geologische Karten als Prognosedatensatz für die hochauflösende Regionalisierung von Bodeneigenschaften verwendet, setzt sich die Grenzziehung der Geologischen Karte durch die Analyse fort und kann in den Prognoseergebnissen zu Artefakten führen (Fehlerfortpflanzung) (vgl. Gessler et al., 1995; Scull et al., 2005).

Die vorliegende Arbeit will Beiträge zur Verbesserung der Prognose von bodenkundlichen Informationen in großen Landschaften im Rahmen der Boden-Landschaftsforschung liefern und neue methodische Ansätze zum Umgang mit großen Datenbeständen und großen Landschaftsräumen aufzeigen.

Die bearbeiteten Schwerpunkte sind Bestandteil der bodenkundlichen Arbeiten im Rahmen des Sonderforschungsbereichs 299 (Landnutzungskonzepte für periphere Regionen) der deutschen Forschungsgemeinschaft (DFG). Übergeordnetes Ziel des SFB 299 ist die Entwicklung und Anwendung einer integrierten Methodik zur Erarbeitung und Bewertung von ökonomisch und ökologisch nachhaltigen, natur- und wirtschaftsräumlich differenzierten Optionen der regionalen Landnutzung (SFB 299, 2005). Bodenkundlich steht die Regionalisierung von Bodeneigenschaften als

Eingangsgröße für hydrologische, ökonomische und ökologische Modelle auf Basis geophysikalischer Messungen im Vordergrund. Untersuchungsgebiet in der vierten Phase des SFB war das Einzugsgebiet der Nidda. Es umfasst weite Teile des Wetteraukreises (Friedberg, Hessen) und hat eine Gesamtausdehnung von 1600 km². Die Landschaft ist dabei geologisch, petrologisch und ökologisch hochdivers und umfasst auch die Mittelgebirgsregionen Taunus, Vogelsberg und den Büdinger Wald (Abb. 1.1).

Für einzelne Verfahrensschritte und Teilmodule sind neben dem Nidda Einzugsgebiet noch Datensätze aus dem Pfälzer Wald, Zentralhessen und der Republik Niger verwendet worden.

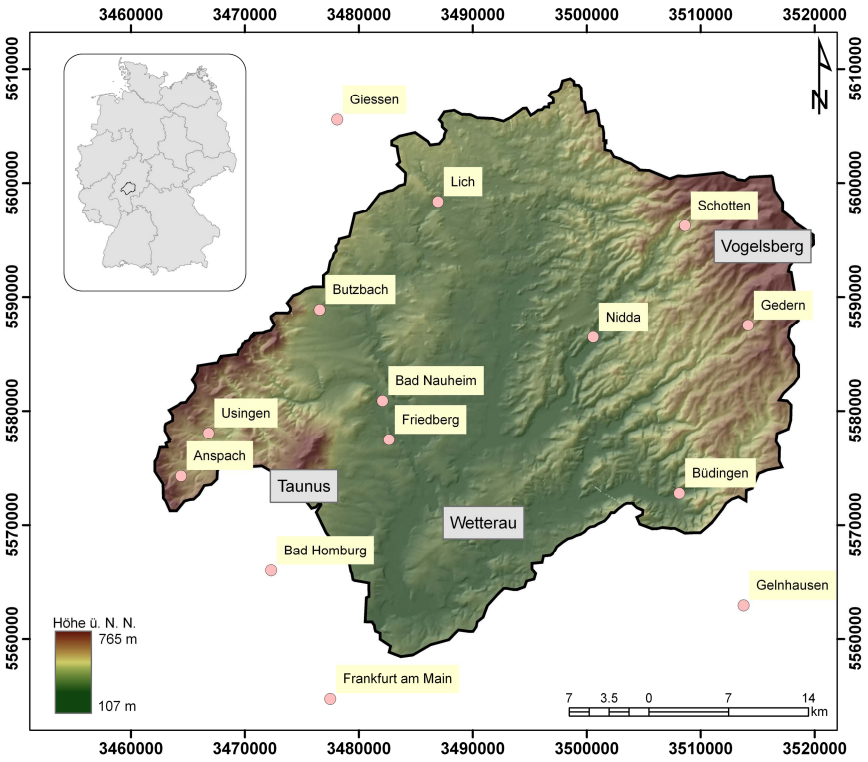


Abb. 1.1: Lage des Nidda Einzugsgebiets und Höhendifferenzierung, das Rheinische Schiefergebirge mit dem Taunus rahmt zusammen mit dem Vogelsberg im Westen die lössbedeckte Wetterau ein.

2 Ausgangssituation und Stand der Forschung

2.1 SFB 299 Teilprojekt B1

Der überwiegende Teil der Arbeiten in dieser Dissertation wurde im Rahmen des Teilprojektes B1 des 1997 an der Justus-Liebig Universität Gießen eingerichteten Sonderforschungsbereichs 299 (Landnutzungskonzepte für periphere Regionen) der DFG durchgeführt. Zentraler Aspekt des Projektbereichs B war die Erfassung und Regionalisierung abiotischer und biotischer Prozessgrößen und Zustandsvariablen, die die ökologischen und ökonomischen Landschaftsfunktionen steuern.

Die beiden bodenkundlichen Teilprojekte B1 und B2.1 befassten sich im Einzelnen mit folgenden Forschungsthemen:

- Räumlich repräsentative Beprobung
  (Szibalski et al., 1999; Szibalski, 2001; Behrens et al., 2001, 2008, 2009a)
- Einsatz des Georadars zur Erfassung von periglaziären Lagen und Substraten
  (Schotte und Felix-Henningsen, 1999; Sauer und Felix-Henningsen, 2004; Gerber et al., 2004; Gerber et al. 2008)
- Entwicklung von Prognoseverfahren und Regionalisierung von Bodeneigenschaften
  (Szibalski, 2001; Scholten et al., 2001; Scholten und Behrens, 2002; Scholten, 2003; Behrens et al., 2005; Behrens und Scholten, 2006a,b; Schmidt et al., 2008, 2009)
- Digitale Reliefanalyse als Basis von Bodenprognosen
  (Behrens, 2003; Behrens et al. 2009b,c)

Die Untersuchungsgebiete in der ersten SFB Phase waren drei Gemarkungen im Lahn-Dill-Bergland. In der zweiten und dritten Projektphase wurde der Untersuchungsraum auf das gesamte Lahn-Dill-Bergland erweitert. In der vierten Phase stand das mit 1600 km^2 doppelt so große Einzugsgebiet der Nidda im Zentrum der Arbeiten.

Durch eine kontinuierliche methodische Weiterentwicklung wurde ein integratives Konzept aus raumstatistischen und bodenkundlichen Methoden und Modellierungstechniken in Kombination mit geophysikalischen Messverfahren entwickelt, das eine hochauflösende Regionalisierung von Bodeneigenschaften auch in großen Landschaften ermöglicht.

Die bodenkundliche Erfassungs- und Regionalisierungsmethodik, wie sie im SFB in Teilprojekt B1 entwickelt und in der letzten Phase komplettiert und angewendet wurde, ist ein mehrstufiges Verfahren. Zu Beginn werden die Basisdaten der jeweiligen Großlandschaft analysiert. Umfasst die Großlandschaft mehrere naturräumliche Einheiten, wird die Landschaft segmentiert, um die so entstandenen homogeneren Landschaftssegmente separat analysieren zu können und so bessere Prognoseergebnisse zu ermöglichen. Die nächsten beiden Schritte dienen der Vorbereitung geophysikalischer Messungen. Zum Einen werden in den Landschaftssegmenten repräsentative Teilräume ausgewiesen, die in ihrer Struktur (Relief- und Bodenverbreitung) den Landschaftssegmenten möglichst ähnlich sind (Behrens et al., 2009a). Damit ist eine Übertragung der Ergebnisse aus den Teilräumen in die Landschaftssegmente sichergestellt. Zum Anderen werden in den ausgewiesenen Teilräumen im nächsten Schritt repräsentative Transekte ausgewiesen, die durch alle, in dem jeweiligen Teilraum vorkommenden Einheiten verlaufen und somit ebenfalls eine Übertragbarkeit gewährleisten. Im Anschluss erfolgt die

hochauflösende Aufnahme der Substrate entlang der repräsentativen Transekte mit dem Georadar (Gerber et al., 2004). Nach der Prozessierung der Georadardaten (Gerber et al., 2008) werden die Ergebnisse anschließend mit Data Mining-Verfahren wie Künstlichen Neuronalen Netzen oder Entscheidungsbaumverfahren regionalisiert (Lagacherie und Holmes, 1997; Zhang et al., 1999; Zhu, 2000; Behrens et al., 2005; Scull et al., 2005; Schmidt et al., 2009).

2.2 Konzepte der Repräsentanz

Unter *Repräsentanz* wird im Allgemeinen die möglichst genaue Abbildung einer Grundgesamtheit anhand einer Stichprobe verstanden. Die Anzahl der Stichproben hängt dabei von der Komplexität der jeweiligen Grundgesamtheit ab (Haseloff und Hoffmann, 1970; Bahrenberg et al., 1990; Jouan-Rimbaud et al., 1997; Schröder et al., 2004). Die *Grundgesamtheit* stellt somit die inhaltliche, zeitliche und/oder räumliche Struktur von Merkmalen dar. Eine *Stichprobe* beschreibt die Grundgesamtheit über eine Teilmenge. Die Entwicklung von repräsentativen Beprobungsstrategien nimmt somit in der Boden-Landschaftsmodellierung eine Schlüsselposition zur Erstellung valider, reproduzierbarer und übertragbarer Prognosen ein.

Grundsätzlich können drei Kategorien von Stichproben unterschieden werden: expertenbasierte, zufallsbasierte oder systematische Stichproben. Je nach angewendetem Verfahren ergeben sich Unterschiede in der Verallgemeinerung bzw. der Übertragbarkeit, also in den Rückschlüssen, die aus der Stichprobe auf die Grundgesamtheit gezogen werden können. Die mit der expertenbasierten Stichprobenauswahl verbundene Integration einer subjektiven Komponente in den Entscheidungsprozess limitiert den wissenschaftlichen Anspruch hinsichtlich Reproduzierbarkeit und Übertragbarkeit (Fränzle, 1978; Kuhnt, 1994; Lagacherie et al., 1995; Schmotz, 1996; etc.). Somit sind rein expertenbasierte Stichproben nicht oder nur eingeschränkt anwendbar. Unabhängige, zufallsbasierte Stichproben vereinen den Vorteil statistischer Validität durch die Möglichkeit der Berechnung von Vertrauensintervallen bzw. Vertrauensniveaus, die als Repräsentanzkriterium dienen können und ermöglichen gleichzeitig die Verallgemeinerung der Verteilungsfunktion auf die Grundgesamtheit. Jedoch werden in vielen Fällen relativ viele Proben benötigt, um mit einer zufallsbasierten Stichprobe den kompletten Merkmalsraum abzudecken.

Systematische Stichprobenverfahren ermöglichen es, bei geringerer Probendichte Bereiche der Grundgesamtheit zu erfassen, die in einer gleichgroßen, zufallsbasierten Stichprobe nicht berücksichtigt würden. Zu solchen Verfahren gehören beispielsweise geschichtete Stichprobenverfahren. Zufallsbasierte Verfahren (random sampling) stehen dabei den gezielten fragestellungsbezogenen Verfahren (purposive sampling) gegenüber. Letztere werden für Fragen des Digital Soil Mapping empfohlen (Brus et al., 2006).

Neben der klassischen Ausweisung von Probenahmestandorten fallen unter den Begriff der Repräsentanz auch komplexere raumstatistische Analysen, die für vielfältige Fragestellungen im Rahmen der Bodenlandschaftsmodellierung von Interesse sind. Dazu zählt beispielsweise die Bewertung der räumlichen Repräsentativität von existierenden kleinräumigen Bodenkarten (Favrot, 1989, zit. in Lagacherie et al. 1995, 2001). Die so genannte „*reference area*"-Methode sucht dabei nach Gebieten, die einem kleinen, detailliert kartierten Gebiet ähnlich sind. Somit lassen sich Gebiete ausweisen, die für eine Extrapolation auf Basis vorliegender Daten in Frage kommen. Solche Verfahren bieten somit die Möglichkeit einer komplexen raumanalytischen, modellbasierten Stichprobenauswahl und –bewertung, die mit reinen designbasierten Zufallsverfahren nicht möglich ist.

Im Rahmen der Arbeiten im SFB 299 wurden unterschiedliche Verfahren zur repräsentativen Auswahl von Probenahmestandorten, Untersuchungsgebieten und Transekten angewendet und entwickelt. In der ersten Projektphase wurde das Verfahren der Regionalen Assoziationsanalyse nach Kuhnt (1994) eingesetzt, um unter Berücksichtigung lokaler Nachbarschaftsverhältnisse räumlich-repräsentative Beprobungspunkte auszuweisen (Szibalski et al., 1999). In der zweiten und dritten Phase des SFB 299 wurde von Behrens et al. (2001, 2009b) eine raumstatistische Methode entwickelt, die es ermöglicht, repräsentative Teilräume auszuweisen. Dazu wird innerhalb einer Landschaft nach einem kleineren Gebiet gesucht, dass eine hohe Ähnlichkeit in Bezug auf Ausstattung und Zusammensetzung mit dem Ursprung besitzt. Dies ist von Bedeutung, wenn der gesamte Untersuchungsraum zu groß ist, um flächendeckend beprobt werden zu können, bzw. wenn Monitoring-Flächen ausgewiesen werden sollen, die dem Großraum möglichst ähnlich sind. Inhaltlich stellt es somit ein zur „*reference area*"-Methode (Lagacherie et al. 1995; 2001) gegenläufiges Verfahren dar, da noch keine detaillierten Informationen zur räumlichen Extrapolation vorliegen.

Sollen durch den Einsatz geophysikalischer Messverfahren wie beispielsweise Georadaraufnahmen Daten zur Extrapolation von Bodeneigenschaften in größere Landschaftsgebiete erhoben werden (Gerber et al., 2008), ist die Ausweisung von repräsentativen Messtransekten erforderlich, um eine Übertragbarkeit der Messungen zu ermöglichen. Auch hier bieten sich systematische, repräsentative Verfahren zur Bestimmung der Lage der Transekte an (Kapitel 3.1.2).

2.3 Bodenprognose

Verfahren und Methoden zur Prognose von Bodentypen bzw. Bodenformen sowie von Bodeneigenschaften sind in der Literatur vielfach beschrieben (McBratney et al., 2000; Scull et al., 2003; Behrens und Scholten, 2006b). Allgemeines Ziel der digitalen Bodenkartierung ist die

Regionalisierung lückenhaft oder punktuell vorliegender Bodendaten. Grundsätzlich können zwei Methodenkomplexe zur räumlichen Beschreibung unterschieden werden: Interpolationsverfahren und Extrapolationsverfahren.

Interpolationsverfahren ermöglichen die flächenhafte Charakterisierung innerhalb der vom vorliegenden Datensatz bestimmten räumlichen Grenzen. Einfache räumliche Interpolationsverfahren, die meist zur Regionalisierung metrisch skalierter Daten verwendet werden, wie die Inverse Distanzgewichtung (IDW, Shepard, 1968; Robinson und Mettemicht, 2006) und Thiessen-Polygone (Thiessen, 1911) stehen dabei komplexeren Verfahren wie dem Kriging (Matheron, 1963) gegenüber.

Im Gegensatz dazu ermöglichen *Extrapolationsverfahren* Regionalisierungen auch außerhalb der räumlichen Grenzen einer Stichprobe. Sie können sowohl auf metrische als auch auf kategorische Daten angewendet werden. Die Regionalisierung erfolgt dabei durch Regressions- oder überwachte Klassifikationsverfahren auf Basis von Sekundärinformationen (unabhängigen Variablen) wie z.B. Reliefparametern. „Überwacht" bedeutet in diesem Zusammenhang, dass Klassen auf Basis von vorliegenden Daten regionalisiert werden. Ergebnisse von unüberwachten Klassifikationsverfahren wie beispielsweise der Clusteranalyse basieren allein auf der statistischen Verteilung der unabhängigen Variablen und müssen immer expertenbasiert interpretiert werden, da sie per se keine bodenkundliche Information darstellen.

Regressions- und Klassifikationsverfahren werden oft unter dem Begriff Data Mining-Verfahren zusammengefasst. Der Einsatz von Data Mining-Verfahren auf bodenkundliche Fragestellungen – insbesondere der Verbreitungssystematik und der Bodengenese - sind Teil der rasant wachsenden bodenkundlichen Forschungszweige der Pedometrie und der digitalen Bodenkartierung („Digital Soil Mapping").

Je nach Herkunft der verschiedenen Regressions- und Klassifikationsverfahren wie z.B. der Informatik, der Mathematik, der Forschung zur künstlichen Intelligenz oder der Statistik werden die Verfahren in die Teildisziplinen Maschinelles Lernen (*Machine Learning*), Erkenntnisgewinnung (*Knowledge Discovery*), statistisches Lernen (*Statistical Learning*) oder computergestütztes Lernen (*Computational Learning*) eingeordnet. Zentrale Verfahren sind u.a. Künstliche Neuronale Netze (Zhu, 2000, Behrens et al., 2005), Entscheidungsbaumverfahren (Lagacherie und Holmes, 1997; Scull et al., 2005, Schmidt et al., 2008) und Support Vector Machines (Behrens und Scholten, 2006b).

Die Anwendung von überwachten Extrapolationsverfahren auf bodenkundliche Fragestellungen ist stets von umweltrelevanten Informationen als Stützvariablen zur räumlichen bzw. inhaltlichen Beschreibung der Zielvariablen abhängig. Die Charakterisierung der wichtigen Faktoren des Prozessgefüges Boden geht auf die Forschungsarbeiten von Dokuchaev (1893) und

Jenny (1941) zurück. Die Integration der einzelnen pedogenetisch wirksamen Faktoren (Klima, Relief, Geologie, Organismen, Zeit) in Extrapolationsmodelle ermöglicht somit eine Modellierung der Zusammenhänge und deren räumliche Abbildung. McBratney et al. (2003) erweitern das von Jenny (1941) formulierte Modell zur Bodengenese speziell für die Anwendung in quantitativen empirischen räumlichen Bodenprognosemodellen. Danach lässt sich eine Bodeneigenschaft (s) nach folgender Funktion berechnen:

$$s = f(s, c, o, r, p, a, n)$$

mit:

s = Bodeneigenschaft
c = Klimaparameter
o = Organismen (Flora, Fauna, Mensch)
r = Reliefparameter
p = Bodenbildendes Substrat
a = Alter/Zeit
n = Räumliche Lage/Nachbarschaft.

Neben der Verfügbarkeit umweltrelevanter Informationen zum Aufbau von *scorpan*-Modellen ist die Untersuchungsgebietsgröße und die räumliche Auflösung der Daten ein wichtiger Aspekt, der die Anwendung von Data Mining-Verfahren limitiert. Dabei nimmt mit steigender Untersuchungsgebietsgröße und steigender Auflösung die Datensatzgröße schnell zu, sodass auch mit aktuellen Workstations (4 Kerne, 8 GB Arbeitsspeicher) die Daten nicht mehr sinnvoll prozessiert werden können.

Moran und Bui (2002), Grinand et al. (2008) und Schmidt et al. (2008) zeigen erste Konzepte zum Umgang mit großen Datenmengen im Bereich von Bodenprognosen. Dies ist unter anderem der Fall, wenn es sich um große Stichprobenmengen handelt, wie sie beispielsweise aus der Rasterung von Bodenkarten in Geographischen Informationssystemen entstehen. Stichprobenverfahren (Kapitel 3.2.1) sind dabei ein wichtiger Schritt zur Handhabung und Nutzbarmachung von Daten sowie zur Verbesserung von Prognoseergebnissen.

Andererseits nimmt die Verfügbarkeit von Prädiktoren (unabhängigen Variablen für die Prognose) zu. Sind bisherige Modellierungsstudien zu räumlich bodenkundlichen Fragestellungen vorwiegend auf Basis weniger umweltrelevanter Parametersätze (< 20) durchgeführt worden, zeigt die Studie von Behrens et al. (2009d), dass durch die effektive Kombination von unterschiedlichen räumlichen Skalenniveaus und angepassten Prognoseverfahren deutliche Verbesserungen in der Prognosequalität erreicht werden können. Die damit einhergehende Zunahme der Dimensionalität in den umweltrelevanten Daten erschwert allerdings den Einsatz von Modellierungsverfahren zur Prognose von

Bodeninformationen im Landschaftsrahmen. Weiterer Forschungsbedarf ist hier notwendig.

Eine Literaturauswertung zum Thema Bodenprognose im Hinblick auf Untersuchungsgebietsgröße und räumliche Auflösung für den Skalenbereich vom Teileinzugsgebiet bis hin zur Großlandschaft offenbart ein deutliches Defizit an Prognosen mit hoher Auflösung in Großlandschaften (Tab. 2.1). Es zeigt sich, dass erst in den letzten drei bis vier Jahren Arbeiten mit Auflösungen < 50 m in Verbindung mit Prognosegebietsgrößen von über 500 km² durchgeführt wurden. Die Arbeiten von Ryan et al. (2000) und Campling et al. (2002) erreichen bei einer Auflösung von 25m und einer Untersuchungsgebietsgröße von über 500 km² eine hohe räumliche als auch inhaltliche Auflösung. Im direkten Vergleich werden jedoch vorwiegend kleinere Gebiete hochauflösend prognostiziert. Jüngere Studien von Behrens et al. (2005), Ziadat (2007), Grinand et al. (2008) und Giasson et al. (2008) zeigen die allgemeine Tendenz hin zu höheren Auflösungen und größeren Gebieten. Dies ist einerseits dem bestehenden Bedarf der hochauflösenden Beschreibung von Landschaften geschuldet, andererseits der steigenden Verfügbarkeit von leistungsstarken Rechnerarchitekturen und effizienten analytischen Methoden.

Tabelle 2.1: Literaturauswertung nach Untersuchungsgebietsgröße und Auflösung der Prognosedaten sortiert nach dem Erscheinungsjahr (Kursiv: Studien die im Rahmen dieser Arbeit entstanden sind).

Autoren	E.-jahr	Zeitschrift	Untersuchungsgebiet in km²	Auflösung in m
Walker et al.	1986	Soil Sci. Soc. Am. J.	0,007	10
Bhatti et al.	1991	Remote Sensing of the Environment	0,26	15
Moore et al.	1993	Soil Sci. Soc. Am. J.	0,054	15
Gessler et al.	1995	Int. J. Geogr. Inf. Sci.	100	10
Skidmore et al.	1996	PE & RS	0,97	10
Cook et al.	1996a	Soil Sci. Soc. Am. J.	0,054	15
Cook et al.	1996b	Australian Journal of Soil Research	200	70
Cialella et al.	1997	PE & RS	24	10
Lagacherie and Holmes	1997	Int. J. Geogr. Inf. Sci.	35	50
Skidmore et al.	1997	Int. J. of Remote Sensing	0,18	10
Sinowski und Auerswald	1999	Geoderma	1,5	12,5
Chaplot et al.	2000	Geoderma	0,02	10
McBratney et al.	2000	Geoderma	0,42	2
Ryan et al.	2000	Forest Ecosystem and Management	500	25
Bishop und McBratney	2001	Geoderma	0,47	5
Thompson et al.	2001	Geoderma	0,13	10
Zhu et al.	2001	Soil Sci Soc Am J.	36	30
Campling et al.	2002	Soil Sci Soc Am J.	589	25
Florinsky et al.	2002	Env. Modeling and Software	0,67	15
Kravchenko et al.	2002	Soil Sci. Soc. Am. J.	0,2	10
Park and Vlek	2002	Geoderma	0,03	10
Peng et al.	2003	Geoderma	0,57	10
Behrens et al.	2005	J. Plant Nutr. S. Sci.	600	20
Thompson and Kolka	2005	Soil Sci. Soc. Am. J.	15	30
Ziadat	2007	Geoderma	148	20
Grinand et al.	2008	Geoderma	900	50
Schmidt et al.	*2008*	*Geoderma*	*350*	*20*
Giasson et al.	2008	DSM with Limited Data, Springer	720	92
Zhu et al.	2008	DSM with Limited Data, Springer	60	10
Penizek und Boruvka	2008	DSM with Limited Data, Springer	83	10
Schmidt	*2009*	*Dissertationsschrift*	*1600*	*20*

2.4 Ziele der Arbeit

In Anlehnung an Teilprojekt B1 des SFB 299 ist das Hauptziel dieser Arbeit die Entwicklung und Anwendung einer Methodik zur hochauflösenden und flächendeckenden Beschreibung von Bodeneigenschaften in Großlandschaften auf Basis von prognostischen Verfahren und geophysikalischen Daten. Diese Arbeit will Beiträge zur qualitativen und quantitativen Verbesserung von Bodenprognosemodellen in großen Gebieten im Rahmen der Bodenlandschaftsmodellierung liefern und methodische Ansätze zum Umgang mit großen Datenbeständen und großen Landschaftsräumen aufzeigen.

Die methodischen Ansätze lassen sich in drei Bereiche gliedern, die Grundlage für eine repräsentative und schrittweise Erfassung bodenkundlicher Parameter und deren Extrapolation in einer Großlandschaft sind:

- Landschaftssegmentierung,
- Repräsentanz und
- Data Mining.

Das Ziel der Landschaftssegmentierung (Manuskript 1) ist die Ausweisung von homogenen, nicht-fragmentierten Teileinzugsgebieten mit statistischen Verfahren die reproduzierbare Ergebnisse liefern. Dabei wird der Einsatz von *moving-window*-basierten Häufigkeitsanalysen in Kombination mit einer *k-means Clusteranalyse* getestet. Die Ergebnisse der Landschaftssegmentierung sollen die Grundlage für die Ausweisung repräsentativer Teilräume und Transekte liefern. Die Teilräume bilden dabei die Verbindung zwischen den Landschaftssegmenten und den Messtransekten. Die Ableitung der Teilräume basiert auf einem Verfahren, dass innerhalb des SFB 299 entwickelt wurde. Die Transekte stellen die Grundlage für eine geophysikalische Aufnahme der Untersuchungsgebiete dar.

Somit sollen nach der Segmentierung und der Ausweisung von Teilräumen repräsentative Transekte nach folgenden Kriterien erfasst werden (Manuskript 2):

- jede Raumeinheitenklasse muss im Transektverlauf vorkommen,
- jede Raumeinheitenklasse darf nur einmal vorkommen und
- die Transekte sollten möglichst kurz sein.

Damit ist sichergestellt, dass die gewonnenen Ergebnisse in die Teilräume und anschließend in die Landschaftssegmente übertragbar sind.

Insbesondere im Landschaftsmaßstab erfordert die hochauflösende Modellierung eine technische Optimierung zur Handhabung der enormen Datenmengen (Manuskript 3). Dies ist insbesondere dann von Bedeutung, wenn, wie im Falle der vorliegenden Bodenkarten, kleinmaßstäbig gerasterte Bodeninformationen prognostiziert werden sollen. Aus diesem Grund

wird eine effektive Integration von Stichprobenverfahren in den Prognoseprozess zur Verbesserung der Prognosequalität getestet. Neben der Handhabung großer Datenmengen ist die Qualität der Datengrundlage für den Prognoseprozess von Bedeutung (Manuskript 4). Um die variable Qualität von Datengrundlagen zu adressieren, die Eingang in Prognosen finden, soll abschließend eine Methodik entwickelt werden, die mit Hilfe raumstatistischer Verfahren die Qualität der Datengrundlage überprüft und Fehler in der Grenzziehung nominaler Flächendaten wie Bodenkarten automatisiert korrigiert.

3 Überblick zu den Veröffentlichungen

3.1 Repräsentanz

3.1.1 Landschaftssegmentierung

(Manuskript 1, J. Plant Nutr. and Soil Sci., submitted in July 2008)

Die Qualität der Anwendung von Methoden und Verfahren der digitalen Bodenkartierung wird in starkem Ausmaß von der jeweiligen Lerngrundlage bestimmt. Dabei spielen zum Einen die Qualität und der Detaillierungsgrad eine wichtige Rolle. Zum Anderen müssen die Daten repräsentativ erhoben worden sein, um eine Generalisierung erreichen zu können. Je höher die Heterogenität einer Landschaft, respektive Bodenlandschaft, ist, umso mehr Proben sind für eine qualitativ hochwertige Bodenlandschaftsmodellierung erforderlich. Somit muss als oberste Ebene in einem Beprobungsverfahren die Analyse der Landschaft stehen. Dabei kann es sich durchaus als sinnvoll erweisen, eine Landschaft zu gliedern und separate Modelle für die jeweiligen Segmente zu entwickeln (McBratney et al., 1991). Dieses Vorgehen entspricht dem pedologischen Ansatz, der als zwingende Voraussetzung zur Bewertung von Bodenressourcen die Kenntnis über die Vergesellschaftungsstrukturen der Böden benötigt (Pullan, 1969). Dieses Konzept der Vergesellschaftung von Böden wird seit dem frühen 20. Jahrhundert diskutiert (z.B. Milne, 1935; Jenny, 1941; Ruhe, 1956; Butler, 1959; Schmidt, 1975). Die bekanntesten Konzepte sind dabei der Catena-Ansatz (Milne, 1935) und die Bodenfunktionsgleichung (Dokuchaev, 1893; Jenny, 1941). Um die enorme Komplexität und Variabilität der Bodenvergesellschaftung in einer Landschaft statistisch zu beschreiben, existieren verschiedene raumstatistische Ansätze und Konzepte (Fränzle, 1978; Kuhnt, 1994; McSweeny et al., 1994; Lagacherie et al, 2001; Behrens et al., 2009c).

Im Rahmen des SFB 299 wird versucht, eine Fläche von ca. 1600 km² bodenkundlich hochauflösend abzubilden. Das Einzugsgebiet der Nidda ist jedoch mit seinen geologisch und petrologisch variablen Mittelgebirgsregionen (Taunus, Vogelsberg, Büdinger Rotliegend-/Buntsandstein-Landschaft) als auch der zentral gelegenen, landwirtschaftlich intensiv genutzten, lössbedeckten Wetterau, äußerst heterogen. Um in einer solch diversen Landschaft sinnvolle Generalisierungen mit Hilfe von Data Mining-Verfahren zu erreichen, empfiehlt es sich, die Landschaft vor der Beprobung zu segmentieren. Dazu wurde ein statistisches Verfahren entwickelt, das nominal skalierte, räumliche Datensätze teilautomatisiert auf Basis der Vergesellschaftungsstrukturen innerhalb der Basisdaten in homogene Segmente gliedert, die im Rahmen einer stufenweisen Beprobungsstrategie einzeln beprobt werden. Das allgemeine Ziel der Segmentierung ist die Ausweisung von homogenen, nicht bzw. gering fragmentierten Untereinheiten mit geglätteten Grenzverläufen innerhalb einer Landschaft. Das Einzugsgebiet der Nidda ist somit ein optimales Untersuchungsgebiet für die Entwicklung, Analyse und

Validierung des methodischen Konzepts.

Die sinnvolle Strukturierung der Landschaft unter dem Gesichtspunkt der Reduzierung der Heterogenität und der semi-automatisierten Erfassung von Vergesellschaftungsmustern in Form von Bodenlandschaften basiert im Rahmen des SFB auf einer vorliegenden Bodenkarte 1:50.000 (HLUG) für das Einzugsgebiet der Nidda. Die Verwendung einer existierenden Bodenkarte im Maßstab 1:50.000 als Grundlage für eine Segmentierung der Landschaft scheint im Hinblick auf eine höher aufgelöste Regionalisierung von Bodeneigenschaften mit einer Auflösung von 20 m als optimale integrale Datengrundlage, da sie per se die raumrelevanten pedogenetischen Faktoren subsumiert. Ist keine Bodenkarte verfügbar, können jedoch auch geomorphometrische und petrographische Informationen aus der Digitalen Reliefanalyse sowie Ableitungen aus Geologischen Karten im Sinne von Raumeinheiten (Kuhnt, 1994) verschnitten werden. Eine Segmentierung kann dann auf Basis dieser pedogenetisch wirksamen Faktorenkombination erfolgen.

Zur Beurteilung der Segmentierungsergebnisse wurden die naturräumliche Gliederung 1:200.000 von Meynen und Schmithüsen (1953) und die geologische Übersichtskarte 1:300.000 (HLUG) herangezogen.

Technisch basiert das Verfahren zur reproduzierbaren Ausweisung von Bodenlandschaften auf der Verknüpfung einer *moving-window*-basierten Häufigkeitsanalyse mit einer *k-means Clusteranalyse*. Dabei wird der nominal skalierte Datensatz mit einem *moving-window* abgetastet und für jeden Ausschnitt die Häufigkeitsverteilung der Merkmale in der Karte registriert. Die Häufigkeiten werden im Anschluss mit Hilfe der *k-means Clusteranalyse* ausgewertet. Durch die Veränderung der Fenstergröße und der Clusteranzahl kann der Zersplitterungsgrad eingestellt werden. Zusätzliche Analysen ermöglichen die teilautomatisierte Optimierung dieser beiden wichtigen Modellparameter. Somit kann in Abhängigkeit der jeweiligen Landschaft eine optimierte Segmentierung erfolgen. Analog der Bezeichnung in der naturräumlichen Gliederung wurden folgende Bodenlandschaften ausgewiesen:

- Taunus,
- Südwestliche Wetterau,
- Nordöstliche Wetterau,
- Unterer/Vorderer Vogelsberg,
- Vogelsberg,
- Büdinger Wald.

Detaillierte Untersuchungen im Hinblick auf die geologische und topographische Homogenität

innerhalb der Bodenlandschaften zeigten eine eindeutige und verifizierbare Trennung hinsichtlich der naturräumlichen Ausstattung im Nidda-Einzugsgebiet. Insbesondere durch die Betrachtung von charakteristischen lokalen und regionalen Reliefparametern zeigt die vorgeschlagene Methodik zur Segmentierung von Bodenlandschaften plausible und reproduzierbare Ergebnisse (Abb. 3.1). Die deutlich stärker geneigte, niedrigere Fläche des Büdinger Waldes hebt sich von der angrenzenden Einheit des Vogelsberg und des Unteren Vogelsberg deutlich ab. Zwischen den Vogelsbergregionen fallen deutliche Unterschiede in den Höhenverhältnissen und der Gewässernetzstruktur auf. Selbst innerhalb der beiden Wetterau-Segmente können Differenzierungen anhand der untersuchten Reliefparameter vorgenommen werden.

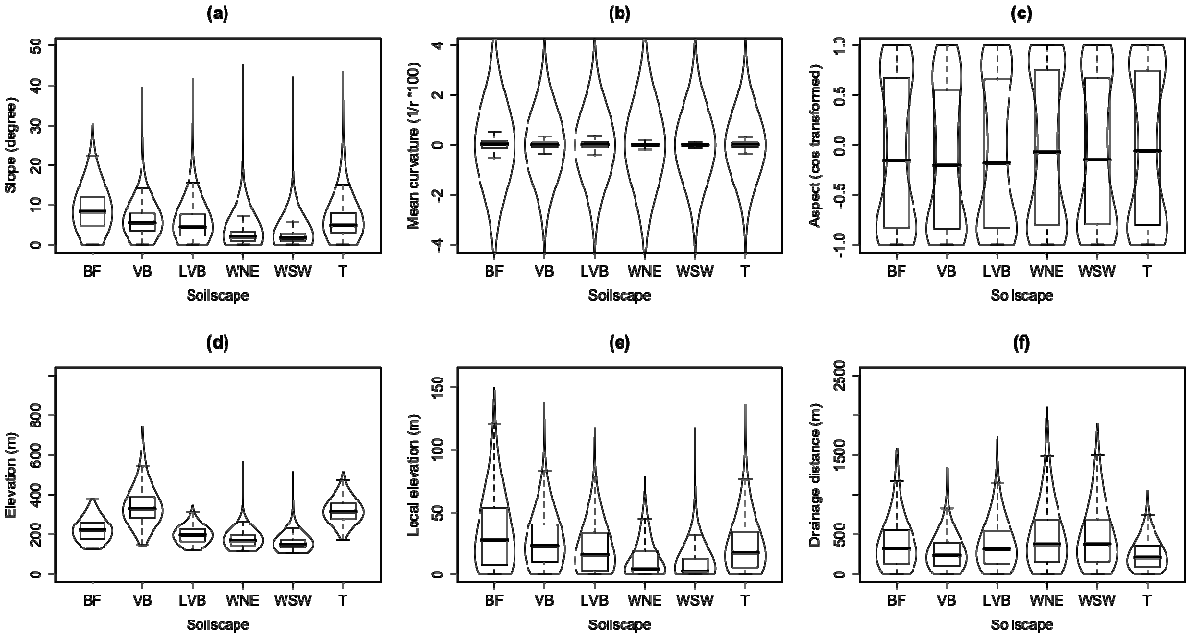


Abb. 3.1: Analyse (Boxplot und Dichtefunktion) charakteristischer lokaler und regionaler topographischer Attribute (slope (a), local elevation (b), aspect (c), drainage distance (d), mean curvature (e)) innerhalb der einzelnen Bodenlandschaftssegmente: Büdinger Wald (BF), Vogelsberg (VB), Unterer Vogelsberg (LVB), südwestliche Wetterau (WSW), nordöstliche Wetterau (WNE) und Taunus (T).

Über eine detaillierte Analyse ausgewählter Reliefparameter hinaus wurden separat die geologischen Verhältnisse in den einzelnen Segmenten untersucht und denen der naturräumlichen Gliederung von Hessen gegenüber gestellt.

Die als Vergleichsreferenz herangezogene naturräumliche Gliederung nach Meynen und Schmithüsen (1953) zeigt aufgrund ihrer deutlich breiter gefassten inhaltlichen Auslegung Schwächen in der detaillierten Beschreibung von Bodenlandschaften, da neben geologischen, petrologischen und geomorphologischen Parametern auch ökologische Parameter wie Artenreichtum und -verteilungen eingeflossen sind. So erreicht die entwickelte Methodik eine höhere Homogenität im Hauptausgangssubstrat wie auch in der individuellen

Vergesellschaftungsstruktur innerhalb der Bodenlandschaften. Darüber hinaus ist die Methodik zum Einen reproduzierbar und zum Anderen auf beliebige Landschaften übertragbar, in denen auf Expertenwissen basierende naturräumliche Gliederungen oder ähnliche landschaftsbeschreibende Kartenmaterialien fehlen.

Ein wesentlicher Forschungsbeitrag dieser Arbeit ist die Kombination verschiedener (raum-) statistischer Verfahren zur Charakterisierung großräumiger Landschaftsstrukturen, wie *moving-window*-basierte Häufigkeitsanalysen, Fragmentationsanalysen und eine *k-means Clusteranalyse*. Neben dem allgemeinen Beitrag zur bodenkundlichen Forschung im Landschaftsrahmen bieten sich fachübergreifende Einsatzmöglichkeiten in der Beschreibung und Diskretisierung von Daten in der ökologischen Forschung (Pennock und Veldcamp, 2006) an. Darüber hinaus kann die Methodik eine allgemeine Grundlage für die hydrologische Modellierung und Kartierung (Sonneveld et al., 2006; Scherrer et al., 2002) wie auch für Modelle zur Beschreibung von Bodenentwicklungsprozessen in Landschaften (Minasny und McBratney, 2006) darstellen.

3.1.2 Repräsentative Transekte

(Manuskript 2, J. Geogr. Inf. Sci., submitted in March 2009)

Die Integration von geophysikalischen Naherkundungsverfahren wie dem Georadar in den Prozess der digitalen Bodenkartierung insbesondere im Bereich der Präzisionslandwirtschaft ist vielfach in Publikationen dargestellt (McBratney et al., 2000, 2003; Viscarra-Rossel et al., 2007). Methodisch-technische Herausforderungen ergeben sich insbesondere bei Anwendungen, die über die Feldskala und insbesondere Nutzungsgrenzen hinausgehen.

Im Rahmen des SFB 299 ist die Landschaftssegmentierung der erste Schritt in einem dreistufigen Beprobungsschema, das zum Ziel hat, kleinräumig aufgenommene Georadarmessungen in große Bodenlandschaften zu übertragen. Nach der Segmentierung der Landschaft in einzelne homogene Bodenlandschaften erfolgt die Ausweisung repräsentativer Teilräume innerhalb der Bodenlandschaften nach einem ebenfalls im Rahmen des SFB 299 entwickelten raum-statistischen Ansatz (Behrens et al., 2009a). Die Ausweisung der repräsentativen Teilflächen innerhalb der Landschaftssegmente (vgl. Kapitel 3.1.1) erfolgt analog auf Basis der Bodenkarte im Maßstab 1:50.000 (HLUG) und einer systematischen Analyse der Ähnlichkeiten in unterschiedlichen Ausschnittsgrößen bezogen auf die ursprüngliche Bodenvergesellschaftung im Landschaftssegment. Die über eine Anpassung des chi^2-Tests ermittelte Ähnlichkeit zum Ursprung definiert die optimale Lage der Untersuchungsfläche im Teilraum (Abb. 3.2).

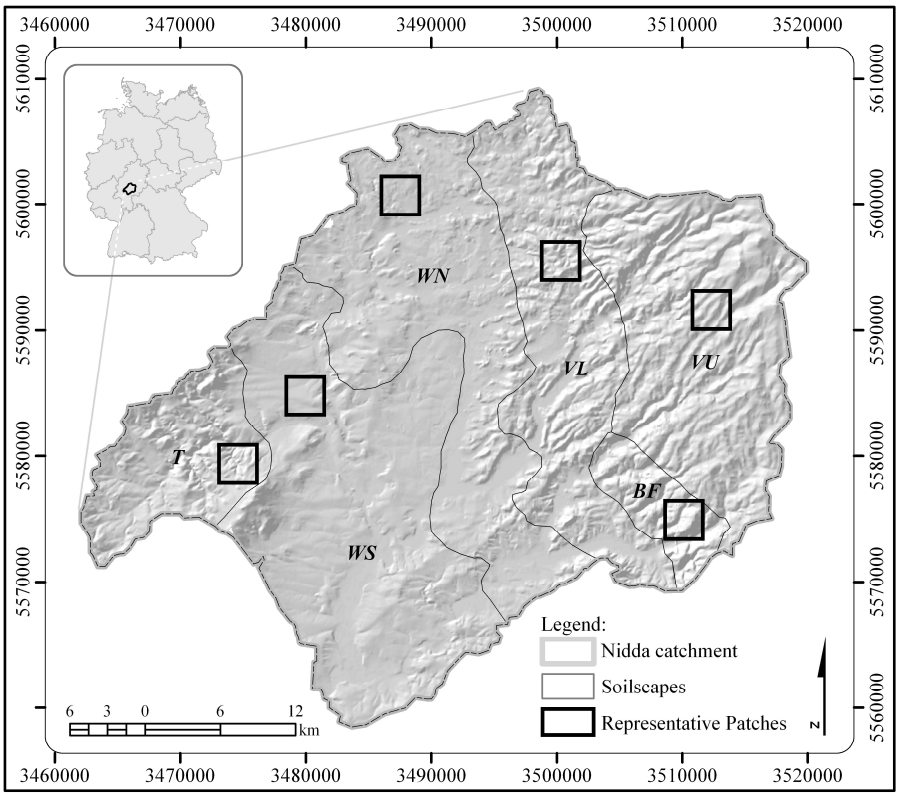


Abb. 3.2: Darstellung der Bodenlandschaftssegmente sowie der räumlichen Lage der einzelnen repräsentativen Teilräume (*BF* = Büdinger Wald; *T* = Taunus; *VU* = Vogelsberg; *VL* = Untere Vogelsberg; *WN* = nord-östliche Wetterau; *WS* = süd-westliche Wetterau).

Die ausgewiesenen repräsentativen Teilräume haben eine Größe von 3x3 km und sind somit klein genug, um als Testflächen für geophysikalische Aufnahmen dienen zu können (Gerber et al., 2008).

Analog zu vielen anderen geophysikalischen Naherkundungsverfahren wird mit dem Georadar kontinuierlich entlang eines linienhaften geophysikalischen Profils gemessen. Im Hinblick auf Fragen der Repräsentanz stellt sich somit die Frage, wo und wie innerhalb des repräsentativen Teilraums gemessen werden soll, um mit Hilfe des Georadars eine valide Lerngrundlage für die anschließende Modellierung zu generieren. Bislang sind in der bodenkundlichen Literatur keine Ansätze zur Beantwortung dieser Fragestellung dokumentiert. Für klassische bodenkundliche Kartierungen in Gebieten ohne Hintergrundinformationen ist ein Verfahren nach Acres et al., (1993) beschrieben, dass sich an Straßenverläufen orientiert und nicht als repräsentativ (und somit übertragbar) eingestuft werden kann.

Basierend auf den vorgestellten Konzepten zur repräsentativen Ausweisung von Bodenlandschaften und Teilräumen auf der Grundlage von nominal-skalierten Daten (hier der Bodenkarte 1:50.000) wurden im Rahmen des SFB 299 zwei Methoden zur repräsentativen

Auswahl von Tansekten für die geophysikalische Erkundung entwickelt: das *Singleline*- und das *Multiline*-Konzept. Die Verfahren sind somit nicht an die Topographie geknüpft, sondern orientieren sich an der räumlichen Verbreitung der Bodenformen.

Die *Singleline*-Methode verbindet alle repräsentativen Bodeneinheiten innerhalb des Untersuchungsgebietes über eine einzelne Linie, wobei jede Klasse per Definition lediglich einmal vorkommen darf, um Redundanzen zu vermeiden. Im Gegensatz dazu verfolgt die *Multiline*-Methodik die repräsentative Beprobung entlang von Übergangsbereichen zwischen einzelnen Bodenformen.

Da Transekte aus Liniensegmenten bestehen und diese wiederum aus Knotenpunkten, wurden aus den nominal-skalierten Flächendaten in einem ersten Schritt Punktdatensätze generiert. Hierzu wurden zwei Verfahren getestet: der Zentrumspunkt-Ansatz und der Zentrumsketten-Ansatz. Als Zentrumspunkt wurde derjenige Punkt innerhalb einer Fläche ausgewiesen, der am weitesten von den Flächengrenzen entfernt liegt. Somit konnte im Gegensatz zum Flächenschwerpunkt (wie er standardmäßig in GIS berechnet wird) sichergestellt werden, dass der jeweilige Punkt nicht zu nah an einer Grenze zwischen zwei Bodenformen und damit im Übergangsbereich liegt. Im Falle des Zentrumsketten-Ansatzes wurde für jede Fläche im Datensatz eine Linie auf Basis eines morphologischen Thinning-Ansatzes (ESRI, 2002) ausgewiesen und in definierten Abständen in Punkte zerlegt (Abb. 3.3).

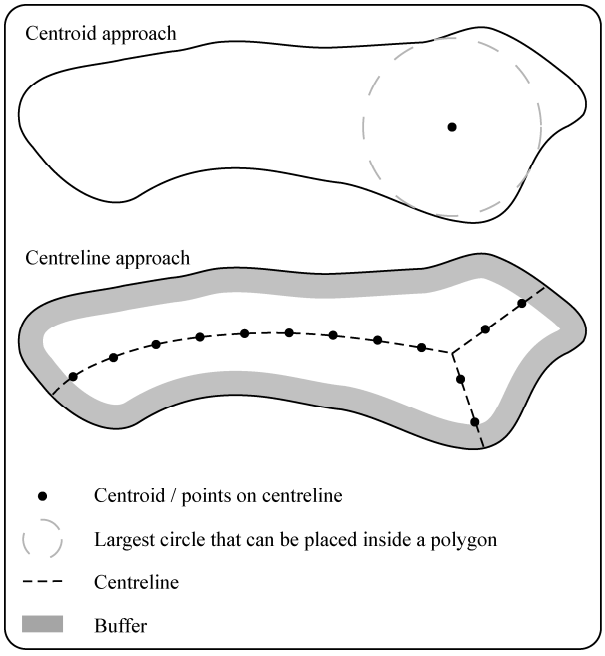


Abb. 3.3: Schematische Darstellung der Arbeitsweise des Zentrumspunkt- und des Zentrumsketten-Ansatzes zur Erzeugung definierter Punktdatensätze.

Zur Herleitung von Transekten auf Basis von Punktdaten bieten sich Verfahren aus dem Bereich der Graphentheorie (Yellen und Gross, 1999) an, wie sie u.a. zur Lösung des „traveling salesman"-Problems verwendet werden. Dabei wird der kürzeste Weg durch eine Punktwolke

gesucht, ohne einen Punkt mehrmals passieren zu müssen. Als Punkte können im Falle eines repräsentativen Transekts die Flächenmittelpunkte der relevanten Raumeinheiten dienen. Im Unterschied zur klassischen Frage des „kürzesten Wegs“ treten im Falle der Ausweisung repräsentativer Transekte folgende Probleme auf:

- es gibt keine vordefinierten Start- und Zielpunkte /-raumeinheiten,
- Raumeinheiten kommen mehr als einmal vor,
- es werden nicht alle Punkte des Datensatzes betrachtet.

Des Weiteren müssen folgenden Kriterien erfüllt sein:

- jede Raumeinheitenklasse muss im Transektverlauf vorkommen
- jede Raumeinheitenklasse darf nur einmal im Transekt vorkommen
- das bzw. die Transekt(e) sollten möglichst kurz sein.

Für den *Singleline*–Ansatz wurde dazu auf Basis eines stochastischen Optimierungsverfahrens (Spall, 2003) ein Suchalgorithmus entwickelt, der über einen definierten räumlichen Nachbarschaftsraum die Verknüpfungspunkte des Transekts identifiziert und sukzessive das Transekt aufbaut. Der *Multiline*-Ansatz basiert auf einer Nachbarschaftsanalyse der Bodenkarte, wobei alle relevanten Nachbarschaftspaare über das kürzeste Transekt auf Basis der abgeleiteten Punkte verbunden wurden. Somit wird im Gegensatz zum *Singleline*-Ansatz eine systematischere Abbildung der Übergangsbereiche erreicht.

Detaillierte Analysen und Modellrechungen an ausgewählten Beispielen zeigen, dass der *Multiline*-Ansatz etwas längere Transekte liefert, jedoch den Merkmalsraum in Bezug auf die Häufigkeitsverteilung charakteristischer Reliefparameter umfassender beschreibt, da er - bei der Anwendung auf eine bereits existierende Bodenkarte - enger mit dem Catena-Konzept nach Milne (1935) verwandt ist.

Der *Multiline*-Ansatz kann darüber hinaus eingesetzt werden, um Übergänge von Bodengesellschaften in Landschaften zu untersuchen und außerdem als Hilfsmittel zur Abschätzung der Qualität von Bodenkarten eingesetzt werden.

3.2 Datenhandling und Bodenprognose

3.2.1 Datenanalyse

(Manuskript 3, Geoderma 146: 138-146)

Mit der steigenden Verfügbarkeit von Geodaten werden jedoch auch die Limitierungen von Prognosemodellen in Hinblick auf die Integration einer möglichst großen Anzahl beschreibender Faktoren als auch einer hohen Stichprobendichte in Prognosemodelle aufgezeigt (Liu und Motoda, 1998; Brighton und Mellish, 2002).

Redundanzen und verrauschte bzw. fehlerhafte Daten können die Effektivität von Prognosemodellen bzw. die Qualität von Prognosen negativ beeinflussen. Dies kann insbesondere dann auftreten, wenn Rasterdaten als Lerngrundlage verwendet werden. In diesem Fall wird jeder Pixel als Stichprobe oder „*Sample*“ verstanden (Bui et al., 1999; Shrestha et al., 2004; Behrens et al., 2005). So können Lerndatensätze entstehen, die mehrere 100.000 „*Samples*“ enthalten.

Die aus dem Forschungsfeld des statistischen Lernens stammenden Verfahren „*Instance Selection*“ (Liu und Motoda, 2001) und „*Feature Selection*“ (John et al., 1994) beschäftigen sich mit der effektiven Reduktion großer Datenmengen. Dabei wird mit Hilfe der Verfahren aus dem Bereich der „*Feature Selection*“ versucht, die Dimensionalität (Anzahl beschreibender Attribute) der Datensätze zu minimieren und relevante Attribute aus dem Gesamtdatensatz zu extrahieren. Im Gegensatz dazu ist es das Ziel der Anwendung von Verfahren aus dem Bereich der „*Instance Selection*“, die Dimensionalität zu erhalten, jedoch redundante oder unscharfe Informationen (*Samples*) aus dem Datensatz zu filtern (Abb. 3.4).

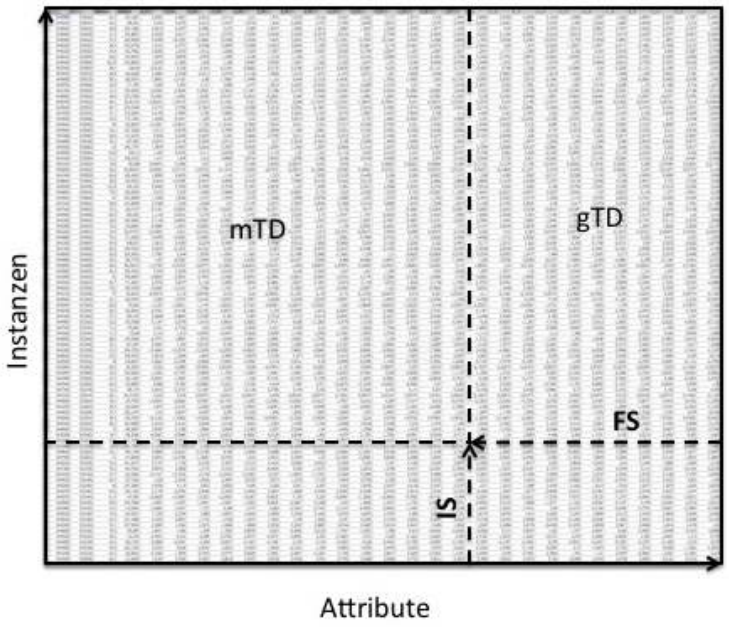


Abb. 3.4: Schematische Darstellung des Einflusses von „*Feature Selection*“ (**FS**) und „*Instance Selection*“ (**IS**) auf den Trainingsdatensatz bezogen auf die Dimensionalität bzw. Stichprobengröße (**gTD** – Gesamtdatensatz, **mTD** – Trainingsdatensatz für die Modellierung).

Die Hauptfunktionen von „*Instance Selection*“-Verfahren werden von Liu und Motoda (2001) als *Enabling*, *Focussing* und *Cleaning* definiert. D.h., dass unter der Prämisse einer zielgerichteten Minimierung eines Lerndatensatzes, bei gleichzeitigem Erhalt der Prognosequalität, die relevanten Informationen herausgefiltert werden. Darüber hinaus wird durch die Reduktion das Verhalten der Prognoseverfahren optimiert.

Grundlage für die Modellierung waren Bodenkarten im Maßstab 1:50.000 aus dem Pfälzer Wald wobei jeweils ein Trainings- und ein Validierungsgebiet (je 40 km²) ausgewiesen wurden. Beide Gebiete verfügen über ein ähnliches Vergesellschaftungsmuster der Bodenformen und über eine ähnliche geomorphologische Raumstruktur (Abb. 3.5).

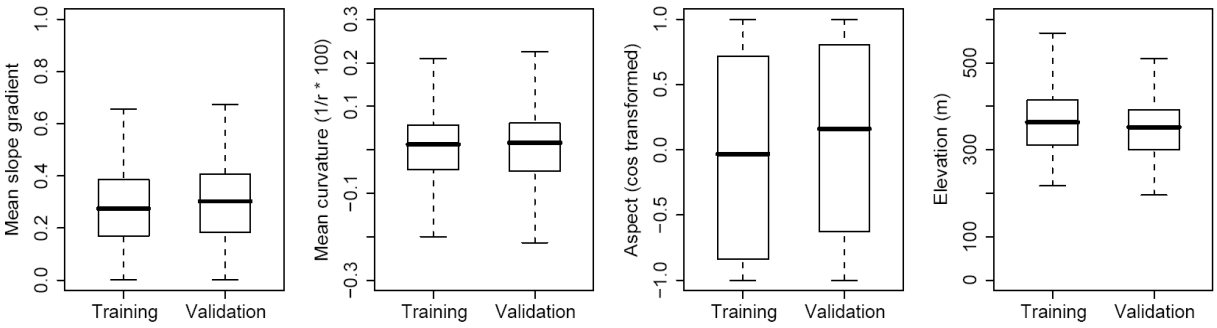


Abb. 3.5: Darstellung des Attributraumes von Hangneigung, Krümmung, Exposition und Höhe über N.N. für das Trainings- und Validierungsgebiet.

In dieser Studie wurden zwei zufallsbasierte „*Instance Selection*“-Verfahren in Kombination mit

einem Entscheidungsbaumverfahren getestet. Der Vorteil von Entscheidungsbaumverfahren liegt in ihrer generellen Robustheit gegenüber hohen Attributdimensionen und Kollinearitäten (Loh und Vanichsetakul, 1988; Lagacherie et al., 2001; Scull et al., 2005). Somit bilden sie eine optimale Grundlage für die systematische Analyse der getesteten Stichprobenverfahren, da der Fokus direkt auf die Stichprobengröße und die damit in Zusammenhang stehende Prognosequalität gelegt werden kann. Zu diesem Zweck wurden die Modellierungsergebnisse aus proportional geschichteten Zufallsstichproben denen disproportional geschichteter Stichproben gegenüber gestellt. Proportional geschichtete Zufallsstichproben beziehen sich dabei auf einen Algorithmus, der die relative proportionale Verteilung der Raumeinheiten innerhalb eines Gesamtdatensatzes auf die definierte Trainingsmenge überträgt. Im Gegensatz dazu verfolgt die disproportional geschichtete Stichprobe eine Gleichgewichtung der einzelnen Raumeinheiten. Die Effektivität der Stichprobenverfahren und die optimale Datensatzgröße wurden anhand der Prognosegüte im Trainings- und Validierungsgebiet bestimmt. Zusätzlich sind unterschiedliche Modellparametereinstellungen des Entscheidungsbaums getestet worden, um auch diesen Einfluss auf die Prognosegüte quantifizieren zu können.

Bezogen auf den Trainingsdatensatz mit 95.000 „*Samples*“ wurden mit jedem Stichprobenverfahren jeweils 500, 1000, 2500, 5000, 7500 und 10.000 „*Samples*“ in drei unabhängigen Wiederholungen gezogen. Methodisch wurden so drei unterschiedliche Ansätze getestet:

- Prognoseergebnisse auf Basis der Modellrechnungen für den Gesamtdatensatz,
- Prognoseergebnisse auf Basis der Modellrechnungen für die proportional geschichteten Stichproben,
- Prognoseergebnisse auf Basis der Modellrechnungen für die disproportional geschichteten Stichproben.

Die Ergebnisse zeigen deutlich die Vorteile der Integration von „*Instance Selection*“-Verfahren in den Prozess der digitalen Bodenkartierung. So wird die Prognosequalität einzelner Modelle durch das systematische Ausdünnen des Trainingsdatensatzes nicht nur erhalten sondern sogar verbessert. Dies ist vorwiegend auf die Prozesse des *Enabling* und des *Focussing* zurückzuführen. Durch die Reduktion des Trainingsdatensatzes wird die Komplexität des Modells herabgesetzt und dadurch das Generalisierungsniveau erhöht. Somit ist die Effektivität des Modells direkt mit der Datensatzgröße verknüpft. Durch die Minimierung von Redundanzen und die Fokussierung auf relevante Informationen ist der Aufbau eines Modells möglich, das durch Optimierung den Entscheidungsraum effektiver charakterisiert.

Basierend auf den erzielten Ergebnissen ist es möglich, die ständig steigende Datengrundlage effektiv zu verwalten und in den Prozess der digitalen Bodenkartierung zu integrieren.

Es empfiehlt sich weitere analytische Verfahren, wie z.B. „*Wilson editing*“ (Wilson, 1972; Behrens et al., 2008), „*latin hypercube sampling*“ (Carre et al., 2007) sowie die mögliche Ableitung von Prototypen (Wai et al., 2001) für weitere Optimierungen hinsichtlich der Prognosegüte in zukünftigen Studien zu überprüfen.

Der Versuch, die inhaltliche Aussagenschärfe der Daten durch eine systematische Reduktion der Trainingsdaten zu erhalten, ist direkt mit der Qualität der Eingangsdaten verbunden. So geht dieser Ansatz davon aus, dass räumlich und inhaltlich genau kartiert wurde. Die Reduktion bezieht sich damit auf die durch die Auflösung im Rasterdatensatz induzierte Redundanz von benachbarten Informationen. Nach einem Zitat von Burrough und McDonnell (1998) - „in practice even the best-drawn maps are not perfect“ – ist die räumliche Aussagenschärfe von bodenkundlichen Kartenmaterialien jedoch nicht immer einheitlich. Daraus erwächst die Forderung nach Algorithmen zur räumlichen und somit inhaltlichen Korrektur von nominal-skalierten digitalen Kartenwerken.

3.2.2 Datenkorrektur

(Manuskript 4, Digital Soil Mapping with limited data. Springer, published in 2008.)

Verfahren der digitalen Bodenkartierung sind in erster Linie von der Qualität der Eingangsdaten abhängig. In vielen Anwendungen kommen jedoch zwangsläufig Daten mit unterschiedlichen räumlichen, zeitlichen und inhaltlichen Auflösungen zusammen und bilden im Rahmen ihrer Integration in Prognosemodellen die Quelle unterschiedlichster Fehler (Robinson et al., 1984; Lagacherie und Holmes, 1997; Heuvelink, 1998; Bishop et al., 2006). Die Datengrundlage nimmt somit die wichtigste Position bezüglich der resultierenden Aussagenschärfe und Interpretierbarkeit ein. Für die Prognose von Bodentypen im Landschaftsmaßstab stellen digitalisierte Bodenkarten im Allgemeinen die wichtigste verfügbare Lerngrundlage dar. Klassische Bodenkarten sind oft nach unterschiedlichen Normen und Kartierungstechniken aufgenommen worden und, da sie expertenbasierte Werke darstellen, meist subjektiven Einflüssen unterworfen (Burrough und McDonnell, 1998; Zhu et al., 2001, Moran and Bui, 2002). Die damit verbundenen Ungenauigkeiten und Generalisierungen zeichnen sich meist unmittelbar in den Grenzverläufen nach (Burnett und Blaschke, 2003).

Im Rahmen von verschiedenen Studien wurde daher ein Verfahren entwickelt, dass die Grenzverläufe von Bodenkarten oder Geologischen Karten mit Unterstützung der digitalen Reliefanalyse anpasst. Die Funktionalität des Verfahrens wurde anhand eines künstlich erzeugten Datensatzes validiert und mit der Anwendung auf Ausschnitte der Bodenkarte

1:50.000 von Hessen und einer Geologische Karte 1:1.000.000 der Republik Niger verifiziert.

Die Korrektur der Grenzverläufe in nominal skalierten Daten erfolgt auf Basis zweier aufeinander folgender Arbeitsschritte im Rasterformat. Im ersten Schritt werden die unsicheren bzw. ungenauen Datenbereiche detektiert und entfernt. Im Anschluss werden die entstandenen Lücken mit Hilfe von räumlichen und nicht-räumlichen Prognoseverfahren geschlossen.

Die Identifikation der Lageungenauigkeit der Grenzverläufe kann zum Einen unüberwacht, basierend auf einem räumlichen Puffer erfolgen und zum Anderen überwacht, durch eine Ausreißeranalyse in den Reliefparametern. Die Ausreißeranalyse erfolgt dabei für jede kartierte Einzelfläche separat. Somit sind detaillierte Reliefinformationen ein entscheidendes Element und nehmen eine übergeordnete Position ein. Für die Analyse wurden 25 Reliefattribute auf Basis einer Reliefanalyse nach Behrens et al. (2008) abgeleitet.

Für eine stabile und nachvollziehbare Klassifizierung des Unsicherheitsbereichs werden neben räumlichen *„Nearest Neighbor“*-Algorithmen unterschiedliche *„Data Mining“*-Methoden, wie Klassifikationsverfahren, *„Feature Selection“* und Filtertechniken angewendet.

Durch die iterative Anwendung der Verfahren ist es möglich, eine plausible und valide Grenzkorrektur vorzunehmen, wobei der Algorithmus die Anpassung unterbricht, sobald in der nächsten Iteration keine signifikanten bzw. plausiblen Änderungen auftreten.

Die Abbildung 3.6 zeigt den iterativen Prozess der Anpassung einer ellipsiodalen Raumeinheit auf Basis eines künstlichen hemisphärischen Geländemodells. Die sichtbaren Ungenauigkeiten im Grenzbereich konnten nach 12 Iterationen entfernt und somit die räumliche Aussagenschärfe verbessert sowie die inhaltliche Genauigkeit konkretisiert werden.

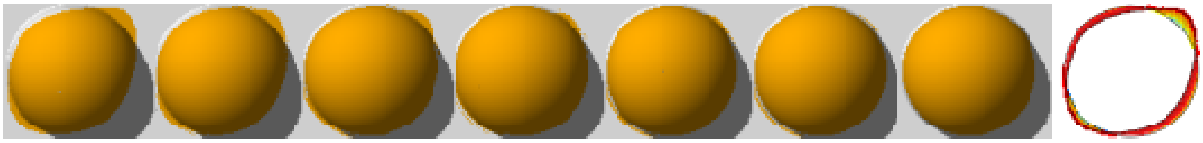


Abb. 3.6: Iterative Korrektur einer ellipsoidalen Raumeinheit auf Basis eines künstlichen hemisphärischen Geländemodells (Links: Ergebnisse nach 2, 4, 6, 8, 10 und 12 Iterationen, Rechts: Lage der korrigierten Pixel).

Die erfolgreiche Kombination der Verfahren im künstlichen Datensatz wurde im nächsten Schritt anhand einer Bodenkarte (1:50.000) im zentralen Teil Hessens überprüft. So zeigen die Ergebnisse dort nur geringe Veränderungen in den Grenzziehungen und sind somit Beweis einer entsprechend hohen räumlichen Genauigkeit. Jedoch konnte mit Hilfe der entwickelten Methodik die Lage einzelner Bodengesellschaften insbesondere in Übergängen zwischen (steilen) Unterhang- und Talbereichen korrigiert werden. Das große Potenzial des vorgestellten Verfahrens liegt insbesondere in der Anwendung in Regionen mit einer geringen Verfügbarkeit

umweltrelevanter Informationen. So zeigt das Beispiel der Korrektur der Geologischen Karte im Maßstab 1:1.000.000 (Greigert, 1961) der Republik Niger anhand eines DGM 90 (SRTM, 2000) die Vorteile der Methodik. In diesem Beispiel wird der zwingende Bedarf an relevanten Reliefinformationen bezogen auf die Eigenschaften der zu korrigierenden Raumeinheit deutlich und erfordert somit eine entsprechend ausgerichtete Analyse bezüglich der Abhängigkeiten der jeweiligen Raumeinheit von den bestimmenden Reliefinformationen. Hier sind u.a. Verfahren wie „*Instance Selection*“, „*Feature Selection*“ und Maschinelles Lernen notwendig, wie sie in den vorangegangenen Kapiteln diskutiert wurden. Daher ist die Komplexität des Ansatzes relativ hoch. Zukünftige Aufgaben betreffen daher die Erweiterung des Verfahrens zur effektiven und zeiteffizienten Attributierung der ungenauen Datenbereiche.

4 Zusammenfassung

Im Rahmen dieser Arbeit, die im Sonderforschungsbereich 299 (Landnutzungskonzepte für periphere Regionen) der deutschen Forschungsgemeinschaft angesiedelt ist, wurden Methoden und Konzepte entwickelt, die eine repräsentative Beprobung und hochauflösende Regionalisierung von Bodeneigenschaften in großen Landschaftsräumen (> 1000 km²) ermöglichen. Der Hauptschwerpunkt lag dabei auf der Entwicklung einer mehrstufigen Beprobungsstrategie in Kombination mit speziellen Verfahren aus dem Bereich des Data Mining zur Regionalisierung von Bodeneigenschaften und Bodenformen. Untersuchungsräume waren das Einzugsgebiet der Nidda in Hessen, der Pfälzer Wald in Rheinland Pfalz und eine Region innerhalb der Republik Niger.

Im Rahmen der mehrstufigen Beprobungsstrategie wurde das Nidda-Einzugsgebiet im ersten Arbeitsschritt (Manuskript 1) in homogene, nicht fragmentierte Teilräume untergliedert. Die als Grundlage für die Segmentierung vorliegende Bodenkarte 1:50.000 (HLUG) wurde mit Hilfe einer *moving-window*-basierten Häufigkeitsanalyse ausgewertet und im Anschluss durch eine *räumliche k-means Clusteranalyse* klassifiziert. Die Anzahl der Cluster, wie auch die Größe des *moving-window*, wurden semi-automatisch ermittelt. Damit konnten die Bodenlandschaften Vogelsberg, Unterer Vogelsberg, Büdinger Wald, Nordöstliche Wetterau, Südwestliche Wetterau und Taunus raumstatistisch voneinander getrennt werden. Somit stehen homogene Bodenlandschaften für bodenkundliche Regionalisierungen zur Verfügung deren Ergebnisse eine hohe Informationsdichte liefern und damit den Beprobungs- und Analyseaufwand minimieren.

Als notwendiger Zwischenschritt zur räumlich repräsentativen Beprobung wurden innerhalb der Bodenlandschaftssegmente repräsentative Teilräume nach einem im SFB 299 entwickelten Verfahren mit einer Größe von 3x3 km ausgewählt.

Im nächsten Arbeitsschritt konnten somit Methoden zur repräsentativen Auswahl von Transekten für die geophysikalische Erkundung entwickelt werden (Manuskript 2). Das *Singleline*-Verfahren erfasst alle Raumeinheiten entlang eines einzelnen Transekts im repräsentativen Teilraum. Die *Multiline*-Methodik weist mehrere Transekte entlang definierter Übergangsbereiche zwischen einzelnen Bodenformen im repräsentativen Teilraum aus. Statistische Analysen und Modellrechungen zeigten, dass der *Multiline*-Ansatz etwas längere Transekte liefert, jedoch den Merkmalsraum in Bezug auf die Häufigkeitsverteilung charakteristischer Reliefparameter wie Hangneigung, Exposition und Höhe über Tiefenlinie umfassender beschreibt.

In einem weiteren Schwerpunkt wird in dieser Arbeit die Eignung von Stichprobenverfahren im Umgang mit redundanten, ungenauen (verrauschten) und großen Lerndatensätzen adressiert,

wie sie bei der Verwendung gerasterter Bodenkarten als Lerngrundlage auftreten können (Manuskript 3). Durch die Anwendung von unterschiedlichen zufallsbasierten Stichprobenverfahren wurde eine Fokussierung auf relevante Informationen ermöglicht und damit die Prognosequalität um bis zu 12% verbessert.

Klassische Bodenkarten wie auch Geologische Karten sind, da sie expertenbasierte Werke darstellen, meist subjektiven Einflüssen unterworfen. Die damit verbundenen Ungenauigkeiten und Generalisierungen zeichnen sich meist unmittelbar in den Grenzverläufen nach. Zur Verbesserung der Grenzverläufe wurde daher im Rahmen dieser Arbeit ein rasterbasiertes Verfahren entwickelt (Manuskript 4). Dabei werden die Unsicherheiten analysiert und an das Relief eines höher aufgelösten Höhenmodells angepasst. Das Verfahren zeigt am Beispiel einer Bodenkarte (1:50.000) im Nidda-Einzugsgebiet sowie einer Geologischen Karte der Republik Niger (1:1.000.000) eine plausible Korrektur der Grenzbereiche, wodurch Fehler und Artefakte in prognostischen Modellen verringert werden können.

Die vorgestellten, modular einsetzbaren Methoden ermöglichen in kombinierter Anwendung eine reproduzierbare und objektive Gliederung der Landschaft in Segmente, eine valide Identifikation repräsentativer Teilräume und Transekte, sowie eine qualitativ verbesserte Geodaten-Basis für die Anwendung in Boden-Landschaftsmodellen. Somit können auch in Großlandschaften valide, hochaufgelöste bodenkundliche Informationen flächendeckend und effizient zur Verfügung gestellt werden.

5 Summary

This thesis is part of the Collaborative Research Centre (SFB) 299 – Land Use Options for Peripheral Regions founded by the German Research Foundation (DFG). The general aim was to develop methods and concepts for representative sampling and high-resolution digital soil mapping in large-scale landscapes (>1000km^2). The main focus was on the development of a multi-stage sampling scheme in combination with data mining techniques in order to regionalize soil properties and soil classes. The main study areas included the Nidda catchment in Hesse, Germany, the Pfälzer Wald in Rhineland-Palatine, Germany, and one region in the Republic of Niger.

As a first step of the sampling scheme (see manuscript 1) the Nidda catchment was divided in homogeneous, non-fragmented segments. The 1:50.000 soil map (HLUG), which served as a basis for the segmentation, was evaluated by a *moving-window* frequency distribution analysis and subsequently classified by a *spatial k-means cluster analysis*. The number of clusters as well as the size of the moving window were determined semi-automatically. Based on this approach the soilscapes Vogelsberg, Lower Vogelsberg, Forest of Büdingen, North-East Wetterau, South-West Wetterau and Taunus could be statistically separated. Thus homogeneous soilscapes for the additional use of digital soil mapping approaches were generated. Based on a patch sampling method developed in the SFB 299 representative patches were derived. The spatial extent of each patch was 3x3 km.

The second step covers the development of methods for a representative generation of transects for geophysical surveys (see manuscript 2). The so called *Singleline* approach includes all spatial units along a single transect in a representative subspace. The *Multiline* method designates several transects along defined transition zones between different soil types in a representative subspace. Statistical analyses and model calculations showed that the *Multiline* approach generates longer transects. With regard to the frequency distribution of characteristic terrain attributes such as slope, aspect and local elevation, however, it describes the feature space more thoroughly.

Further, this thesis focused on an objective and reproducible improvement of the prediction quality of soil information. Traditional soil maps have been compiled with different standards and mapping techniques and, as expert-based values, they are generally affected by subjective influences. The resulting uncertainties and generalizations have a direct effect on the boundaries. Therefore, a raster-based approach was developed in order to adjust the boundaries (see manuscript 3) on the basis of higher resolution terrain attributes. This method produced a plausible adjustment of the boundary regions for a soil map of the Nidda catchment (1: 50.000) and a geological map of the Republic of Niger (1:1.000.000). As a result, errors and artefacts in

predictive models can be reduced.

The last part of this work focused on the combination of sampling schemes and predictive models and their application in order to handle redundant, uncertain (noisy) information in large datasets, as they tend to occur when existing, rasterized soil maps are used as a training base (see manuscript 4). By applying different randomized sampling procedures focussing on relevant information and improving prediction quality up to 12% was enabled.

Based on the modular structure of the methods introduced, it is possible to obtain a reproducible and objective segmentation of the landscape, a valid identification of representative subsets and transects and an improved quality of the geodata basis used for soil-landscape-modelling. Therefore, comprehensive and valid high-resolution soil scientific information can be provided even for large scale landscapes.

6 Verwendete Literatur

Acres, B. D., Green M. A., Rackham, L. J., 1993: A method for identifying soil catenas and determining map unit composition used in a reconnaissance soil survey in Tanzania. Geoderma, 57, pp. 387-404.

Bahrenberg, G., Giese, E., Nipper, J., 1990: Statistische Methoden in der Geographie. Univariate und bivariate Statistik, Teubner, Stuttgart, p. 234.

Behrens, T., 2003: Digitale Reliefanalyse als Basis von Boden-Landschafts-Modellen – Am Beispiel der Modellierung periglaziärer Lagen im Ostharz. Boden und Landschaft, 42, 190p.

Behrens, T., Förster, H. Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005: Digital soil mapping using artificial neural networks. J. Plant Nutr. and Soil Sci. 168, pp. 21-33.

Behrens, T., Purtauf, T., Wolters, V., Köhler, W., Dauber, J., 2001: Study Site Selection and Gradient Detection in Complex Landscapes Using an Automated Patch Detection Tool (PaDS). In: Proceedings of the International Workshop on GeoSpatial Knowledge Processing for Natural Resources Management, pp. 217-221.

Behrens, T., Schmidt, K., Gerber, R., Albrecht, C., Felix-Henningsen, P., Scholten, T., 2009a: Concepts for generating shortest representative transects – sampling approaches for linear operated proximal soil sensors. J. Georg. Inf. Science (submitted).

Behrens, T. Schmidt, K., Scholten, T., 2008: An approach to removing uncertainties in nominal environmental covariates and soil class maps. In: Hartemink, A., McBratney, A., Mendoca-Santos, M.L.,: Digital Soil Mapping with Limited Data. Springer.

Behrens, T., Schmidt, K., Zhu,A.-X., Scholten, T., 2009b: Topography revisited - The ConMap approach for terrain based digital soil mapping. EJS (submitted).

Behrens, T., Schneider, O., Lösel, G., Scholten, T., Hennings, V., Felix-Henningsen, P., Hartwich, R. (2009c): Analysis on pedodiversity and spatial subset representativity – the German soil map 1:1.000.000. J. Plant Nutr. and Soil Sci. (in press).

Behrens, T., Scholten, T., 2006a: Digital Soil Mapping in Germany – a review. J. Plant Nutr. and Soil Sci. 169, pp. 434 - 443.

Behrens, T., Scholten, T., 2006b: Chapter 25. A comparison of data-mining techniques in predictive soil mapping. In Lagacherie, P. McBratney, A.B., Voltz, M. (Eds): Digital Soil mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam. pp. 353-364.

Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2009d: Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma (submitted).

Bhatti, A.U., Mulla, D.J., Frazier, B.E., 1991: Estimation of soil properties and wheat yields on complex eroded hills using geostatistics and thematic mapper images. Remote Sensing of the Enwironment, 37, pp. 181-191.

Bishop, T.F.A., McBratney, A.B., 2001: A comparison of prediction methods for creation of field-extend soil property maps. Geoderma, 103, pp. 149-160.

Bishop, T.F.A., Minasny, B., McBratney, A.B., 2006: Uncertainty analysis for soil-terrain models. Int. J. Geographical Inf. Sci. 20, pp. 117-134.

Brighton, H., Mellish, C., 2002. Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery 6, pp. 153-172.

Brus, D.J., de Gruijter, J.J., van Groenigen, J.W., 2006: Designing spatial coverage samples using the k-means clustering algorithm. In: Lagacherie, P., McBratney, A., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, vol. 3., Elsevier, Amsterdam.

Bui, E.N., Loughhead, A., and Corner, R., 1999: Extracting soil-landscape rules from previous soil surveys. Australian J. of Soil Research, 37, pp. 495-508.

Burnett, C., Blaschke, T., 2003: A multi-scale segmentation/object relationship modeling methodology for landscape analysis. Ecological modeling, 168, pp. 233-249.

Burrough, P.A., McDonnell, R.A., 1998: Principles of Geographical Information System. 2nd ed., Oxford University Press, 356 pp.

Butler, B.E., 1959: Periodic phenomena in landscapes as a basis for soil studies. Commonwealth scientific and industrial research organization, Australia soil publication 14, 20 pp.

Campling, P., Gobin, A., Feyen, J., 2002: Logistic modeling to spatially predict the probability of soil drainage classes. SSSAJ, 66, 1390-1401.

Carre, F., McBratney, A.B., Minasny, B., 2007: Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma, 141, pp. 1- 14.

Chaplot, V., Walter, C., Curmi, P., 2000: Improving soil hydromorphy prediction according to DEM resolution and available pedological data. Geoderma, 97, pp. 405-422.

Cialella, A.T., Dubayah, R., Lawrence, W., Levine, E., 1997: Predicting soil drainage class using remotely sensed and digital elevation data. Photogrammetric Engineering and Remote Sensing, 63, pp. 171-178.

Corwin, D.L., Lesch, S.M., 2005: Apparent soil electrical conductivity measurements in agriculture. Computers and Electronics in Agriculture, 46, pp. 11-43.

Cook, S.E., Corner, R., Grealish, G.J., Gessler, P.E., Chatres,C.J., 1996a: A rule-based system to map soil properties. Soil Science Society of Ameriva, 60, pp. 1893-1900.

Cook, S.E., Corner, R., Groves, P.R., Grealish, G.J., 1996b: Use of airborne gamma radiometric data for soil mapping. Australian Journal of Soil Research, 34, pp. 183-194.

Dokuchaev, V.V. 1893. The RussianSteppes: Study of the Soil in Russia, Its Past and Present. St. Petersburg,Russia: Department of Agriculture Ministry of Crown Domains for the World'sColumbian Exposition at Chicago.

Environmental Systems Research Institute (ESRI), 2002: ArcView 3.3.

Florinsky, I.V., Eilers, R.G., 2002: Prediction of the soil organic carbon content at micro-, meso- and makroscales by digital terrain modeling. 7th World Congress of Soil Science, Bangkok, Thailand, August 14-21, Paper no. 2331.

Fränzle, O., 1978: The Structure of Soil Associations and Cenozoic Morphogeny in Southeast Afrika, In Nagel, H. (ed.): Beiträge zur Quartär- und Landschaftsforschung - Festschrift zum 60. Geburtstag von Julius Fink. Wien, pp. 159-176.

Gerber, R., Felix-Henningsen , P., Behrens, T., Scholten, T., 2008. Applicability of Ground Penetrating Radar as a tool for non-destructive soil depth mapping on Pleistocene Periglacial Slope Deposits. J. Plant Nutr. Soil Sci., (in press).

Gerber, R., Salat, C., Felix-Henningsen, P., Junge, A., 2004: Investigation of the GPR reflection pattern for shallow depths on a test site. Proceeding of the Tenth International Conference on ground penetrating radar. Delft, The Netherlands, 1, pp. 275-278.

Gessler, P.E., Moore, I.D., McKensie, N.J., Ryan, P.J., 1995: Soil-landscape modeling and spatial prediction of soil attributes. Int. J. Geographical Inf. Sci., 9, pp. 421-432.

Giasson, E., Figueiredo, S.R., Tornquist, C.G., Clarke, R.T., 2008: Digital Soil Mapping Using Logistic Regression on Terrain Parameters for Several Ecological Regions in Southern Brazil. In Hartemink, A.E., McBratney, A., Mendoca-Santos (ed.): Digital Soil Mapping with Limited Data. Springer, 445p.

Greigert, J., 1961: République du Niger. Carte géologique de reconnaissance du Bassin des Iullemeden 1:1 Mio. BRGM, Niamey, Niger.

Grinand, C., Arrouays, D., Laroche, B., Martin, M.P., 2008: Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma, 143: 180-190.

Haseloff, O.W., Hoffmann, H.J., 1970: Kleines Handbuch der Statistik. Berlin.

Heuvelink, G.B.M., 1998: Error Propagation in Environmental Modelling with GIS. Taylor &

Francis, London. 144p.

HLUG (Hessische Landesamt für Umwelt und Geologie), 2002: Erläuterungen zur Bodenkarte von Hessen 1:50000. HLUG, 575p.

Jenny, H. (1941): Factors of soil formation. McGraw-Hill: New York, 281p.

John, G. H., Kohavi, R., Pfleger, K., 1994: Irrelevant features and the subset selection problem. Proceedings of the 11th International Conference on Machine Learning, pp. 121-129.

Jouan-Rimbaud, D., Massart, D.L., Saby, C.A., Puel, C., 1997: Characterisation of the representativity of selected sets of samples in multivariate calibration and pattern recognition. Analytica Chimica Acta, 350: pp 149-161.

Kravchenko, A.N., Bollero, G.A., Omonde, R.A., Bullock, D.G., 2002: Quantitative mapping of soil drainage classes using topographical data and soil electrical conductivity. SSSAJ, 66, pp. 235-243.

Kuhnt, G., 1994: Regionale Repräsentanz. Beiträge zu einer Raumorientierten Messtheorie. Habilitationsschrift, Universität Kiel.

Lagacherie, P., Holmes, S., 1997: Addressing geographical data errors in a classification tree soil unit prediction. Int. J. Geogr. Inf. Science, 11, pp. 183-198.

Lagacherie, P., Legros, J.,P., Burfough, P., A., 1995: A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma, 65, pp. 283-301.

Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N., Barthes, J.P., 2001: Mapping of reference area representativity using a mathematical soilscape distance. Geoderma, 101, pp. 105-118.

Lagacherie, P., McBratney, A., Voltz, M., 2006: Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science. Elsevier.

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Norwell, MA. 244p.

Liu, H., Motoda, H., 2001. Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, 448p.

Loh, W.Y. and Vanichsetakul, N., 1988: Tree-structured classification via generalized discriminant analysis. Journal of the American Statistical Association, 83, pp. 715-728.

Matheron, G., 1963: Principles of geostatistics. Economical Geology, 58, 1246-1266.

McBratney, A.B., Minasny, B. & Viscarra Rossel, R. 2006. Spectral soil analysis and inference systems: a powerful combination for solving the soil data crisis. Geoderma, 136, 272–278.

McBratney, A.B., Hart, G.A., McGarry, D. 1991: The use of region partitioning to improve the

representation of geostatistically mapped soil attributes. Journal of Soil Science 42, pp. 513-532.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003: On digital soil mapping. Geoderma, 117, pp. 3-52.

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W., 2002: From pedotransfer functions to soil inference systems. Geoderma 109, 41-73.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbara, M.S., Shatara, T.M., 2000: An overview of pedometric techniques for use in soil survey. Geoderma, 97, pp. 293-327.

McSweeney, K., Slater, B.K., Hammer, R.D., Bell, J.C., Gessler, P.E., Petersen, G.W., 1994: Towards a new framework for modeling the soil-landscape continuum. In Amundson, R., Harden, J., Singer, M.: Factors of soil formation: A fifthieth anniversary retrospective. SSSA Special Publication, 33, pp. 127-145.

Meynen, E., Schmithüsen, J., 1953: Handbuch der naturräumlichen Gliederung Deutschlands. Gemeinschaftsveröffentlichung des Instituts für Landeskunde und des Deutschen Instituts für Länderkunde. Bad Godesberg.

Milne, G., 1935: Some suggested units of classification and mapping, particularly for East African soils. Soil Res., 3, pp. 183-198.

Minasny, B., McBratney, A.B., 2006: Mechanistic soil-landscape modelling as an approach to developing pedogenetic classifications. Geoderma, 133, pp. 138-149.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Petersen, G.A., 1993: Soil attribute prediction using terrain analysis. SSSAJ, 57, pp. 443-452.

Moran, J.C., Bui, E.N., 2002: Spatial data mining for enhanced soil map modeling. Int. J. Geogr. Inf. Sci., 16, pp. 533-549.

Park, S.J., Vlek, L.G., 2002: Environmental correlation of three-dimensional soil spatial variability: a comparison of three adaptive techniques. Geoderma, 109, pp. 117-140.

Peng, W., Wheeler, D.B., Bell, J.C., Krusemark, M.G., 2003: Delineating patterns of soil drainage class on bare soils using remote sensing. Geoderma, 115, pp. 261-279.

Penizek, V., Boruvka, L., 2008: The Digital Terrain Model as a Tool for Improved Delineation of Alluvial Soils. In Hartemink, A.E., McBratney, A., Mendoca-Santos (ed.): Digital Soil Mapping with Limited Data. Springer, 445 p.

Pennock, D.J., Veldkamp, A., 2006: Advances in landscape-scale soil research. Geoderma, 133, pp. 1-5.

Pullan, R.A., 1969: The soil resources of West Africa. in Thomas, M.F., Whittington, G.W. (Editors), Environment and Land Use in Africa. Methuen, London, pp. 147-191.

Robinson, A.H., Sale, R.D., Morrison, J.L. and Muehrcke, P.C., 1984, Elements of Cartography, 5th Edition, John Wiley & Sons, New York, NY, 688p.

Robinson, T.P., Mettemicht, G., 2006: Testing the performance of spatial interpolation techniques for mapping soil properties. Computers and electronics in agriculture, 50, pp. 97-108.

Ruhe, R.V., 1956: Geomorphic surfaces and the nature of soils. Soil Science, 82, pp. 441-455.

Ryan, P.J., McKenzie, N.J., O´Connell, D., Loughhead, A.N., Leppert, P.M., Jacquier, D., Ashton, L., 2000: Integrating forest soils information across scales: spatial prediction of soil properties under Australian forest. Forest Ecology and Management, 138, pp. 139-157.

Sauer, D., Felix-Henningsen, P., 2004: Application of ground penetraiting radar for the determination of the thickness of Pleistocene periglacial slope deposits (PPSD) on hard rocks. J. Plant Nutr. Soil Science, 176, pp. 752-760.

Scherrer, S., Demuth, N., Meuser, A., 2002: A procedure for the identification of dominant runoff processes by field investigations to delineate the relevant contributing areas for flood modelling. In Spreafico, M. and Weingartner, R.: Int. Conf. On Flood Estimation. Proceedings CHR Report II-17.

Schmidt, K., Behrens, T., Scholten, T., 2008: Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma, 146, pp. 138-146.

Schmidt, K., Behrens, T., Scholten, T., 2009: A new method to derive soilscapes by segmenting soil maps for digital soil sensing and mapping. J. Plant Nutr. Soil Science (accepted with revision).

Schmidt, R., 1975: Grundlagen der Mittelmaßstäbigen Landwirtschaftlichen Standort-kartierung. In: Arch. Acker- u. Pflanzenbau u. Bodenkd., Berlin 19, 8, pp. 533-543.

Schmotz, W., 1996: Entwicklung und Optimierung von Verfahren zur flächenhaften Erfassung der Schadstoffgehalte in Böden. IN: EcoSys Beiträge zur Ökosystemforschung, Suppl.Bd.17.

Scholten, T., 2003: Verbreitungssystematik und Eigenschaften pleistozäner periglaziärer Lagen in deutschen Mittelgebirgen. Relief, Boden, Paläoklima 19, 154p.

Scholten, T., Behrens, T., 2002: GIS-gestützte Modellierung der räumlichen Verbreitung und Ausprägung periglaziärer Lagen in Mittelgebirgsregionen. Berichte zur deutschen Landeskunde. 76, pp. 151-168.

Scholten, T., Szibalksi, M., Behrens, T., Felix-Henningsen, P., 2001: Identification of Study Areas and Inter- and Extrapolation of Rwa Data for Pedological and Hydrological Needs. In: King, L., Metzler, M., Jiang Tong (eds.): Flood Risks and Land Use Conflicts in the Yangtze Catchment,

China, and the Rhine River, Germany - Strategies for a Sustainable Flood Management. Schriften zur Internationalen Entwicklungs- und Umweltforschung 2: 205-210. Peter Lang (Europäischer Verlag der Wissenschaften). Frankfurt.

Schotte, M., Felix-Henningsen, P., 1999: Anwendung des Georadars zur Erhebung der Verbreitung und Eigenschaften periglaziärer Lagen im Lahn-Dill-Bergland. - Z. f. Kulturtechnik u. Landentwicklung 40, 220-227. (B 1 B 2.1).

Schröder, W., Pesch, R., Schmidt, G., 2004: Soil monitoring in Germany: spatial representativity and methodical comparability. J. of Soils and Sediments, 4, pp. 49-58.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003: Predictive soil mapping: a review. Progress in Physical Geography, 27, pp. 171–197.

Scull, P., Franklin, J., Chadwick, O.A., 2005: The application of classification tree analysis to soil type prediction in a dessert landscape. Ecological Modelling, 181, pp. 1-15.

SFB 299, 2005: Sonderforschungsbereich 299 der Deutschen Forschungsgemeinschaft: Landnutzungskonzepte für periphere Regionen. Arbeits- und Ergebnisbericht 2003-2005. Justus-Liebieg-Universität Giessen, Giessen.

Shrestha, D.P., Zinck, J.A., Van Ranst, E., 2004: Modelling land degradation in the Nepalese Himalaya. Catena, pp. 135-156.

Sinowski, W., Auerswald, K., 1999: Using relief parameters in a discriminant analysis to stratify geological areas with different spatial variability of soil properties. Geoderma, 89, pp. 113-128.

Skidmore, A.K., Varekamp,C., Wilson, L., Knowles, E., Delaney, J., 1997: Remote sensing of soils in a eucalypt forest environment. Int. J. of Remote Sensing, 18, pp. 39-56.

Skidmore,A.K., Watford, F., Luckananurug, P., Ryan, P.J., 1996: An operational GIS expert system for mapping forest soils. PE & RS, 62, pp. 501-511.

Sonneveld, M.P.W., Schoorl, J.M., Veldkamp, S., 2006: Mapping hydrological pathways of phosphorus transfer in apparently homogeneous landscapes using a high-resolution DEM. Geoderma, 133, pp. 32-42.

Spall, J. C., 2003: Introduction to Stochastic Search and Optimization. Wiley, 618p.

Shuttle Radar Topography Mission (SRTM), 2000: A Mission to Map the World. http://www2.jpl.nasa.gov/srtm/index.html

Szibalski, M., 2001: Großmaßstäbige Regionalisierung labiler Bodenwerte in standörtlich hochdiversen Kulturlandschaften. Boden und Landschaft, 33, 185p.

Szibalski, M., Behrens, T.; Felix-Henningsen, P., 1999: Regionalisierung bodenkundlicher Kennwerte peripherer Regionen am Beispiel des pH-Wertes. - Z. für Kulturtechnik und

Landentwicklung 40, 228-233. (B 1 B 2.1).

Thiessen. A.H., 1911: Precipitation averages for large areas. Monthly Weather Review, 39(7): 1082-1084.

Thompson, J.A., Bell, J.C., Butler, C.A., 2001: Digital elevation models resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma, 100, 67-89.

Thompson, J.A., Kolka, R.K., 2005: Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. SSSAJ, 69, pp. 1086-1093.

Viscarra-Rossel, R., Taylor, J.A., McBratney, A.B., 2007. Multivariate calibration of hyperspectral gamma-ray energy spectra for proximal soil sensing. European Journal of Soil Science, 58, pp. 343-353.

Wai, L., Keung, C.-K., Ling, C.X., 2001: Learning via prototype generation and filtering. In Liu, H. and Motoda, H., 2001 (eds): Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, Boston, 448p.

Walker, P.H., Hall, G.F., Protz, R., 1986: Relation between landform parameters and soil properties. Soil Science Society of America Proceedings, 32: 101-104.

Wilson, D.L., 1972: Asymptotic properties of nearest neighbour rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics 2, pp. 408-420.

Yellen, J., Gross, J.L., 1999: Graph Theory and Ist Applications. CRC Press, 270p.

Zhang, J., Guo, D., Wan, Q., 1999: Geospatial Data Mining and Knowledge Discovery using Decision Tree Algorithm - A Case Study of Soil Data Set of the Yellow River Delta. - Geoinformatics and Socioinformatics, Proceedings of Geoinformatics'99 Conference, Ann Arbor, Michigan, pp. 1-8.

Zhu, A.X., 2000: Mapping soil landscape as spatial continua: The neural network approach. - Water Resources Research 36, pp. 663-677.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001: Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. SSSAJ, 65, pp.1463 - 1472.

Zhu, A.X., Yang, L., Li, B., Qin, C., English, E., Burt, J.E., Zhou, C., 2008: Purposive Sampling for Digital Soil Mapping for Areas with Limited Data. In Hartemink, A.E., McBratney, A., Mendoca-Santos (ed.): Digital Soil Mapping with Limited Data. Springer, 445 p.

Ziadat, F.M., 2007: Land suitability classification using different sources of information: Soil maps and predicted soil attributes in Jordan. Geoderma, 140, pp. 73-80.

Manuskript 1

Generation of soilscapes by segmenting soil maps for digital soil sensing and mapping in homogeneous feature spaces

Journal of Plant Nutrition and Soil Science submitted in July 2008

Karsten Schmidt[1], Thorsten Behrens[1], Klaus Friedrich[2], Thomas Scholten[1]

[1] Institute of Geography, Chair of Physical Geography, Eberhard Karls University Tübingen, Rümelinstraße 19-23, D-72074, Tübingen, Germany

[2] Hessian State Office for Environment and Geology (HLUG), Rheingaustraße 186, D-65203, Wiesbaden, Germany

Abstract

Digital soil mapping for large areas is challenging if mapping resolution should be as high as possible and sampling should be as sparse as possible. Generally, the more diverse a landscape is, the more samples are required to systematically cover the entire feature space. Moreover, if soil sensing approaches like ground penetrating radar are used in a combined soil sensing and mapping approach in large areas it is important to systematically segment the landscape and to derive representative sensing sites.

Segmenting a landscape as introduced in this study is the first part of a stacked sampling scheme to collect representative soil data for digital soil property mapping, developed in the Collaborative Research Centre (SFB) 299 of the German Research Foundation (DFG). It is followed by deriving representative patches and transects for linear operated soil sensing techniques. We introduce a semi-automated method to segment nominal spatial datasets based on the local spatial frequency distribution of the mapping units aiming to provide homogeneous and non-fragmented segments with smoothed boundaries. The methodological framework for segmentation comprises different spatial and non-spatial techniques and is mainly focussing on a moving window analysis of the frequency distribution and a k-means cluster analysis.

Based on an existing soil map at a scale of 1:50.000 in the highly diverse Nidda catchment, Hesse, Germany, comprising an area of 1600 km^2, we derived six segments and compared these with a map of landscape units (1:200.000), comprising eight main landscape units within the catchment. Comparisons with respect to the distribution of soils and parent materials reveal that the proposed approach returns spatial segments with a higher homogeneity in terms of feature space. Similar results were obtained by analyzing the feature spaces of different terrain attributes.

As segmentation is based on a soil map, soilscapes are derived. These can not only be used for sampling purposes, but are of importance for a variety of environmental issues such as biodiversity and ecosystem analyses or characterization of hydrological units.

1 Introduction

The combination of soil sampling, soil sensing, and soil mapping techniques is a challenging key technology, aiming to map soil properties, functions, and threats, on resolutions as high as possible, for areas as large as possible, and with the highest accuracies possible (Werban et al., 2008). As soil sensing techniques like ground penetrating radar cannot be operated across large landscapes entirely (Gerber et al., 2007) the question arises where to operate the sensing techniques to achieve accurate and transferable extrapolation results.

Based on successful applications of treating soilscapes independently as a basis for adapted digital soil mapping approaches (McBratney et al., 1991) it seems reasonable to segment the landscape into soilscapes prior to mapping. Thus, it is important to develop concepts and techniques for automatically generating soilscapes – or landscape segments in general – as pedometrical tools.

The need to segment a landscape into soilscapes as a basis for digital soil mapping stems from the tremendous complexity of soils in landscapes (McSweeny et al., 1994). Thus, McSweeny et al. (1994) proposed to set up a hierarchical multi-stage strategy to explain the variability of soils and soil properties in space. This follows the concept of pedology pointing to the fact that soil resource assessments cannot be made unless the patterns of associations of soils are known (Pullan, 1969). This issue is recognized in pedology since the early 20th century (e.g. Milne, 1935; Jenny, 1941; Ruhe, 1956; Butler, 1959; Schmidt, 1975). Following this concept, the first aim in mapping larger areas should be to account for soil association patterns as a basis to segment landscapes.

One data source providing landscape segments in Germany is the map of landscape units (Naturräumliche Gliederung) available at a scale of 1:200.000 (Meynen and Schmithüsen, 1953), an approach comparable to land systems mapping introduced by Stewart and Perry (1953). Landscape units divide the landscape on the basis of geological, geomorphological, and ecological settings. Thus, using the map of landscapes units seems to be an obvious choice in terms of stratifying the landscape for soil sampling.

However, concerning soil sampling the delineation of the map of landscape units might not be appropriate as soil spatial pattern and other ecological characteristics used for mapping such as species number and their spatial distribution might vary differently (Meynen and Schmithüsen, 1953). Thus, for segmenting landscapes as a basis for soil sensing in soil mapping studies other

approaches should be considered.

One approach might be a landform classification based on digital elevation models (DEM) as described by many authors (cf. Kundert, 1988; Dikau, 1992; Friedrich, 1996; Möller et al., 2008). Concerning landform classification, these terrain-based approaches are only appropriate for finer scales and/or hierarchical dimensions as they mainly focus on deriving terrain facets instead of segmenting larger homogeneous terrainscapes from one another. Therefore, we examine the possibility to segment a landscape on the basis of an existing soil map 1:50.000, comprising all factors of soil formation, to derive soilscapes, a term introduced by Buol et al. (1973) and conceptually extended by Hole (1978) in the context of pedology, pointing to areas with a homogeneous composition of soils. This seems most appropriate for further subsequent soil sampling and soil mapping schemes aiming at higher resolutions and scales as provided by the soil map used for segmentation. Technically, we apply a moving window based analysis of the local soil class frequency distribution with cluster analysis approaches. In terms of feasibility, e.g. for ground penetrating radar sensing along transects, these segments should be spatially non-fragmented and consist of smooth boundaries.

The approach presented here was developed as a first step of a stacked or multi-stage sampling approach for digital soil property sensing and mapping based on GPR surveys within the Collaborative Research Centre (SFB) 299 of the German Research Foundation (DFG). The entire sampling scheme consists of three independent steps: first, the delineation of soilscapes as introduced in this study, which aims at deriving different homogeneous soilscapes for separate surveys. Second, the detection of representative patches (Behrens et al., 2008a) within these units to reduce the investigation area while preserving the information content, and third, the computation of representative transects for surveys with linear operated soil sensing techniques (Behrens et al., 2008b). Following, soil sensing spatial data mining approaches can be applied for mapping (Schmidt et al., 2008).

We compare the soilscapes derived with the landscape units of the existing map of landscape units in terms of soil class distribution, the distribution of parent material based on a geological map, as well as the feature spaces of different terrain attributes.

2 Study area

The study area - the Nidda catchment - comprises about 1600 km^2 and is located in the German mid-latitude landscapes in Hesse, north of Frankfurt/Main (cf. Fig. 1). Due to its large extent it covers landscapes with different geological, geomorphological, and ecological settings such as the Taunus as part of the Rhenish Massif, parts of the Vogelsberg, the largest basalt formation in Central Europe, the Wetterau loess landscape, and the Büdinger Forest characterised by Triassic

bunter beds.

The soil classes found can mainly be described by Cambisols on different substrates, covering about 27 % of the Nidda catchment, as well as Luvisols and relictic Chernozems with a coverage of about 11 % in association with Stagnosols.

The average precipitation varies between 400 and 1000 mm (HLUG, 2008) due to the differences in altitude of about 600 meters.

The slightly undulating Wetterau, an extension of the Upper Rhine Graben, is very fertile and thus one of the oldest cultural landscapes of Germany (Kalis et al., 2003), whereas the surrounding steeper landscapes are dominated by forest usage.

3 Materials and methods

3.1 Datasets

As basis to segment the landscape of the Nidda catchment we used a soil map 1:50.000 of Hesse, Germany (BFD50).

The map of landscape units (Naturräumliche Gliederung) (1: 200.000) of Germany (Meynen and Schmithüsen, 1953) provides landscape units with a uniform distribution of ecological, geomorphological, and geological settings. It is used as a reference to evaluate the soilscapes derived. Therefore, we compare it with the segmentation approach as introduced in this study in terms of environmental feature spaces such as relief, parent material, and soil class distribution. The mapping approach is based on a hierarchical system of orders to allow interpretations of local, regional, and countrywide differences. We use the 4th order of the hierarchical system, comprising eight units within the Nidda catchment (cf. Fig. 1) as reference. In contrast, the lower orders are too coarse in terms of a reasonable segmentation and the finest (5th) order comprises about five times as much units. Thus, they will not be considered for separate soil sampling schemes.

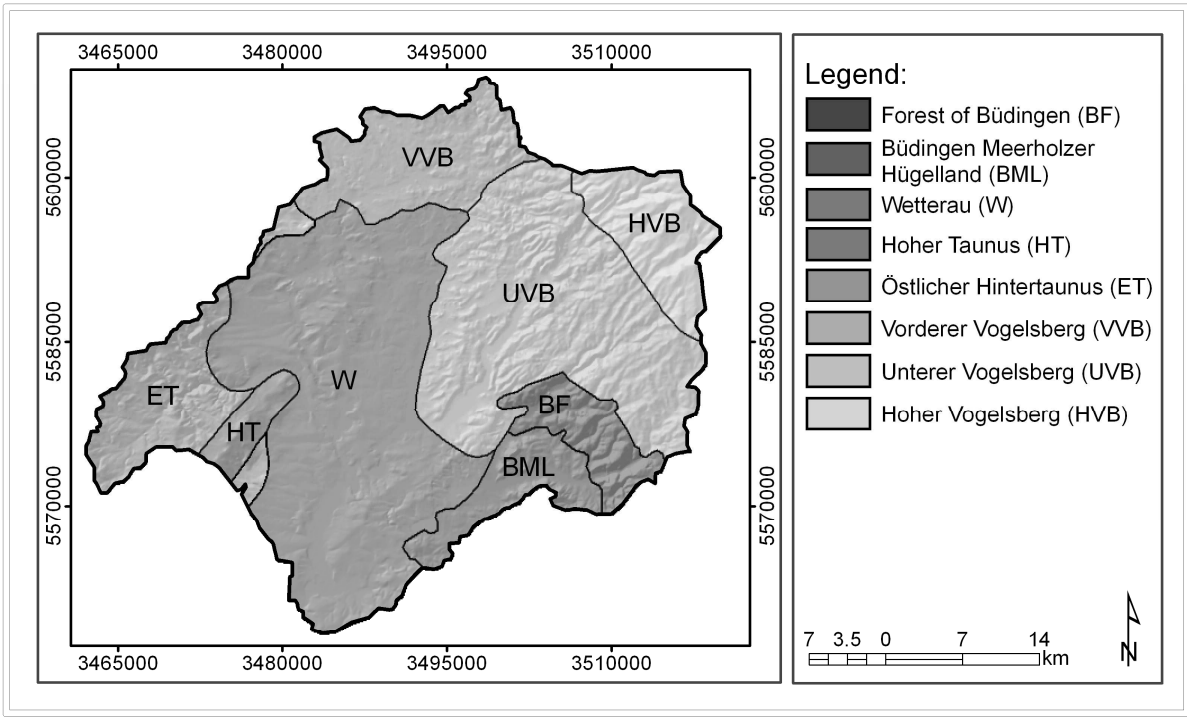


Fig. 1. Map of landscape units in the Nidda catchment (1600km², Hesse, Germany; Klausing, 1988).

Based on the concept of subsequent high resolution digital soil mapping approaches mainly focussing on terrain attributes (McBratney et al., 2003; Behrens et al., 2008c) the feature space of different terrain attributes derived from a DEM with a resolution of 20 m is analyzed to identify and describe the characteristics of each soilscape. We derived slope (Tarboton, 1997), cosine transformed aspect (Horn, 1981), mean curvature (Zevenbergen and Thorne, 1987), and local elevation (Behrens et al., 2008c), which are easily to interpret and classically used as indicators for mapping soil distributions. Additionally, we compared the distance to drainage channels.

Parent material distribution as derived from a geological map (1:300.000) serves as an independent dataset for analyzing the soilscapes derived.

3.2 Spatial segmentation

Following the aim of automatically deriving homogeneous non-fragmented soilscapes with smooth boundaries for digital soil sensing and mapping purposes, different spatial and non-spatial techniques have to be applied. The approach demonstrated in this study is mainly based on a moving window analysis of the frequency distribution of the soil classes and a k-means cluster analysis. Additionally, the optimized window size has to be found which determines the

smoothness of the boundaries as well as the degree of spatial fragmentation.

3.2.1 Local frequency distribution analysis

The local frequency distribution (spatial arrangement/spatial association) of soil classes is the crucial element of the landscape segmentation approach. Therefore, a moving window approach was implemented to compute the frequencies of all soil classes of a map within each window analyzed. Each soil class frequency is then used as a feature in a cluster analysis. We tested window sizes of 250 m, 500 m, 1000 m, 2000 m, …, and 10.000 m.

3.2.2 Spatial k-means cluster analysis

K-means cluster analysis (MacQueen, 1967) is a well-known approach of unsupervised learning. The aim is to compute groups with high interclass variance and small intraclass variance (Webster and Beckett, 1968; Everitt, 1980). The number of clusters has to be defined by the user.

To find an optimum number of soilscapes we tested 3 to 10 classes for each window size. This was done in relation to the eight classes found in the map of landscape units within the Nidda catchment. More than 10 classes had not been tested in terms of feasibility.

In combination with the fragmentation analysis (cf. Chapter 3.2.3), which returns the optimized widow size, we derived the optimized number of classes by calculating the mean perimeter over the soilscapes.

3.2.3 Fragmentation analysis

Using small window sizes apparently returns a more fragmented shape than larger window sizes, as more local noise or hot spots will be averaged out with larger window sizes and more homogenous regions will be returned. Thus, fragmentation analysis is a key element in the approach introduced.

We use the change in the overall perimeter of the soilscapes derived as the criterion to determine the optimal window size and thus the fragmentation. As the perimeter decreases with larger window sizes the optimum window size is found when it stops changing significantly even if the window size is further enlarged. Therefore, an interpretation of the window size vs. perimeter curve is required.

4 Results and discussion

4.1 Spatial cluster analysis

The analysis of the local spatial association of soils reveals an optimum of six soilscapes. The size of the window was set to 5 * 5 km based on the results of the fragmentation analysis. As the objective of this study was to develop a (semi-)automated approach, we chose perimeter measures to evaluate and quantify the optimized number of soilscapes and the optimized window size (cf. Chapter 2.2 and 2.3). Fig. 2 (a) shows the results for finding the optimized window size and Fig. 2 (b) for the number of classes respectively. It can be seen that for window sizes >= 4000 m the mean perimeter is not further increasing significantly and thus window sizes between 4000 m and 6000 m seem to be appropriate. Based on the fragmentation we chose a final window size of 5000 m. Examples for window sizes of 250 * 250 m, 2000 * 2000 m, and 5000 * 5000 m are given in Table 1.

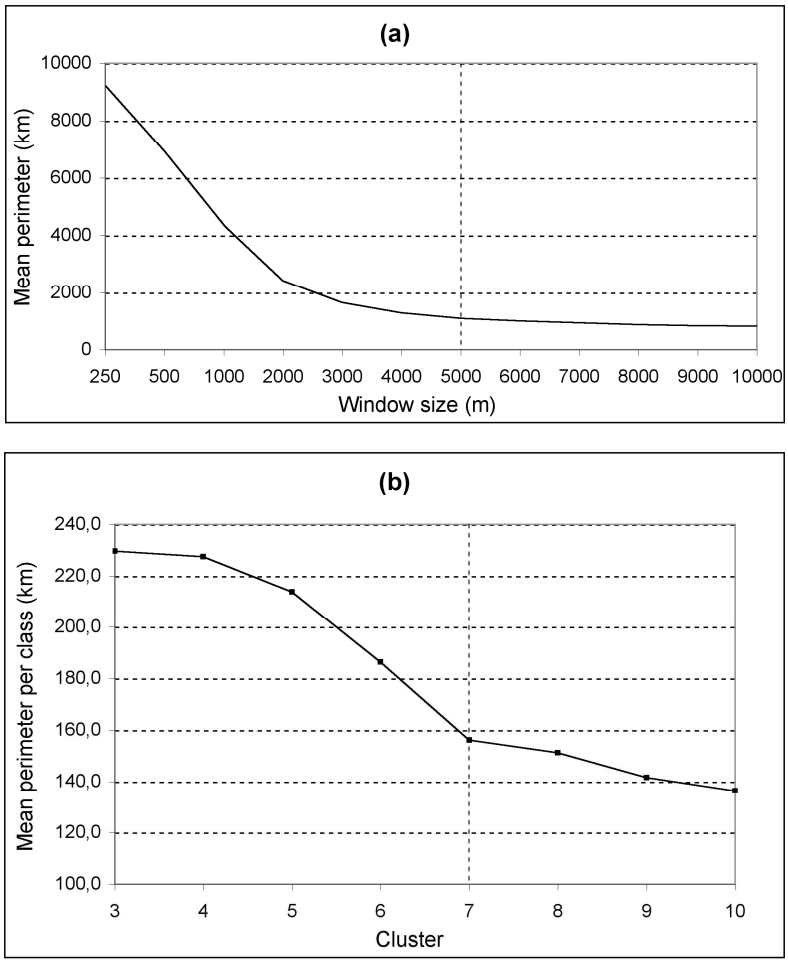


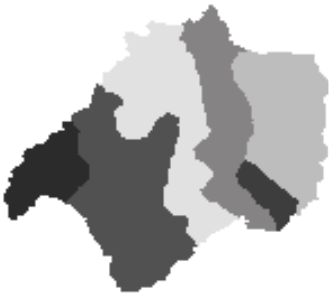
Fig. 2. (a) Fragmentation analysis based on the mean perimeter per window size resulting in an optimum window size of 5000 m (dashed grey line) and (b) optimal number of clusters based on the mean perimeter per class for a window size of 5000 m (dashed grey line).

The analysis for the optimum number of classes by means of the perimeter per class returned 7 clusters. From 3 up to 7 classes the mean perimeter per class decreases strongly. From 7 classes onward this decrease is diminishing, indicating that the soilscapes become more complex in

shape. As the aim of this study is to provide a reasonable number of homogenous and non-fragmented soilscapes we chose 7 classes for the further analysis (cf. Table 1).

Table 1

Different classification results based on window sizes of 250 m, 2000 m and the final map with the optimal window size of 5000 m and the corresponding fragmentation level.

Spatial distribution			
Window size	250 * 250 m	2000 * 2000 m	5000 * 5000 m
Class areas count	1008	92	6

Due to the moving window based analysis approach the pixels located at the boundary of the Nidda catchment contain No Data values. Consequently, one of the 7 classes covers the entire boundary region and was therefore removed by means of Euclidian allocation. The final map of soilscapes as shown in Fig. 3 contains 6 classes.

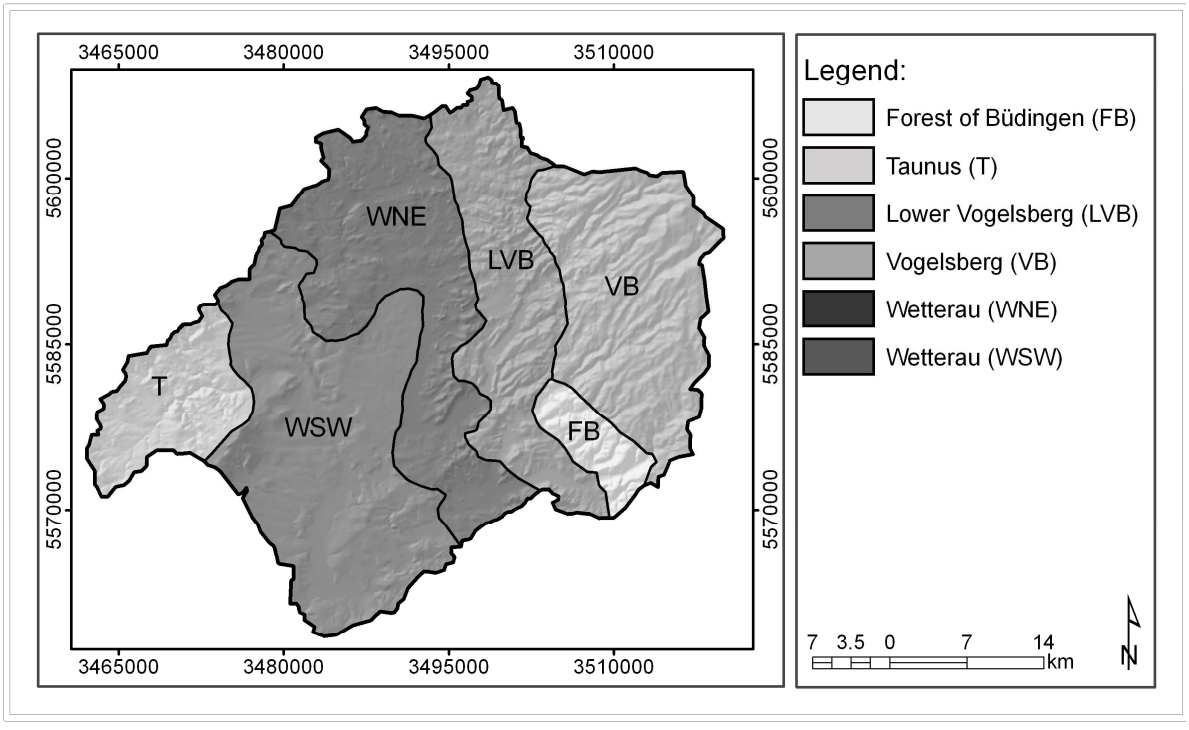


Fig. 3. Soilscapes of the Nidda catchment based on the soil map 1:50.000.

Even though not all boundaries of the derived soilscapes can directly be related to the units of the map of landscape units it is possible to assign landscape names to the six soilscapes of the Nidda catchment, which are: the Taunus (*T*), the Lower Vogelsberg (*LVB*), the Vogelsberg (*VB*), the Forest of Büdingen (*FB*) and the two soilscapes of the Wetterau (*WSW* and *WNE*) [**].

4.2 Analysis of the soilscapes

4.2.1 Soils

Based on the different classification steps with 3, 4, 5, and 6 resulting soilcapes (clusters, Fig. 4) the Vogelsberg (*VB*) differs from the other soilscapes as it is separated with almost the same delineation in all steps.

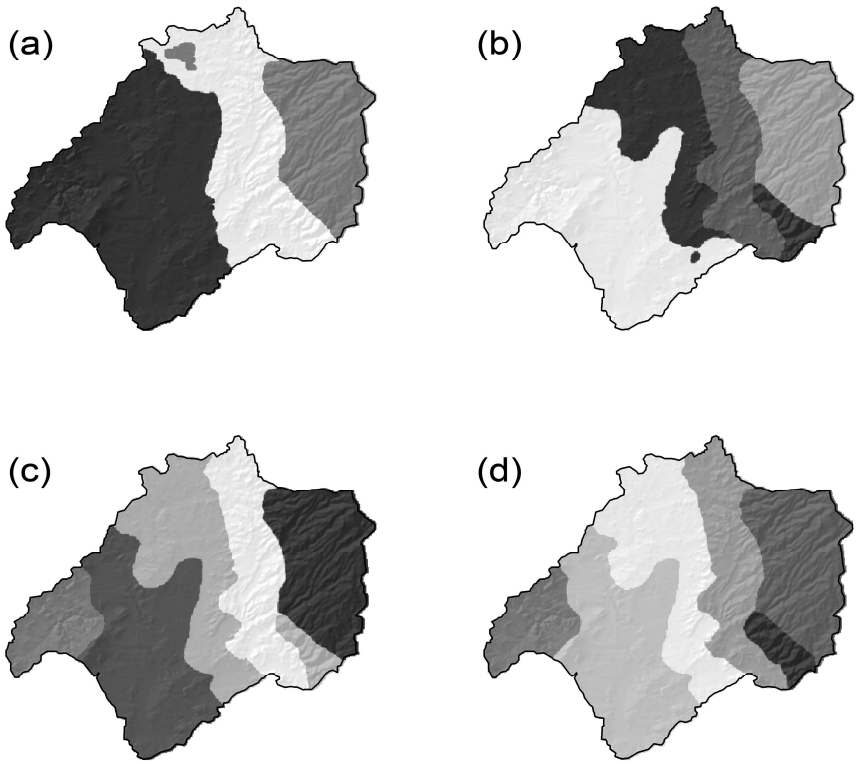


Fig. 3.4. Segmentation results for 3 (a), 4 (b), 5 (c) and 6 (d) clusters

The predominant soils found in the Vogelsberg (*VB*) are Cambisols covering about 38 % of the entire soilscape and Stagnic Luvisols with a coverage of 19 % both formed on Pleistocene Periglacial Slope Deposits (PSD) consisting of Loess intermixed with weathering products of Basanite.

[**] To avoid confusion with the naming schemes of the soilscapes and the landscape units the soilscapes are abbreviated with *italic* letters, whereas the landscape units are abbreviated with regular letters (cf. Chapter 5.3). Additionally, the full names are always provided.

The soilscape following in the west – the Lower Vogelsberg (*LVB*) – is characterized by an almost exact inversion of the frequencies of the soils found in the Vogelsberg (*VB*) (33 % of Stagnic Luvisols and 17% of Cambisols). Furthermore, the percentage of colluvial soils rises from 0.8 % to 9 %. Both soilscapes can be found in almost all classification approaches.

With an increasing number of soilscapes (n > 4; cf. Fig. 4) in the classification approach the Wetterau (*WSW / WNE*) differs from its surroundings. In the southwest part the Wetterau (*WSW*), which is "the Wetterau" in the strict sense, is dominated by soil associations formed on Loess (> 50 %) such as Humic Luvisols (17 %) and (relictic) Chernozems (11 %). The northeast Wetterau (*WNE*) holds an intermediate position between the southwest Wetterau (WSW) and the Lower Vogelsberg (*LVB*). The soilscape is characterized by soils formed on Loess as well as on PSD consisting of weathering products from Loess and Basanite, pointing to higher erosion rates compared to the southwest Wetterau (*WSW*).

The Taunus (*T*) is located in the western part of the Nidda catchment. It is dominated by Cambisols (42 %) associated with stagnic Luvisols (13 %) formed on Slate and Loess.

The Forest of Büdingen (*FB*) – located between the Higher and the Lower Vogelsberg (*LVB*) as well as the Wetterau – consists of stagnic Luvisols (24 %) and (podsolic) Cambisols (22 %) formed on substrates from bunter.

The segmentation of the Nidda catchment into soilscapes is thus mainly based on differences in parent material and the spatial distribution of Loess and PSD.

4.2.2 Feature spaces analysis

The feature space analysis provides further details about the soilscapes derived as well as information about differences in the geomorphological settings, helping to interpret the purpose of the segmentation for a stacked sampling approach in digital soil sensing and mapping mainly based on attributes from digital terrain analysis.

Fig. 5 shows box-plots and kernel density functions of the distribution of terrain attributes for each soilscape. It can be seen that the Forest of Büdingen (*FB*) has the strongest relief characterized by the steepest slopes, the highest differences in elevation, and the widest range of curvature.

The Vogelsberg (*VB*) shows an intermediate relief in terms of steepness, local elevation, drainage network density and curvature. A spatial anisotropy in aspect due to the radial nature of the volcano is revealed.

The generally weakest relief is found in the Wetterau (*WSW / WNE*). The north-eastern part is slightly steeper and has a larger curvature bandwidth. Therefore, the potential for erosion is

higher, so that consequently less loess can be found at present (cf. Chapter 2.1).

Together with the Vogelberg (*VB*) the Taunus (*T*) has the highest drainage density due to their Tertiary and Devonian nature compared to the other soilscapes influenced by Pleistocene Loess accumulations.

For all terrain attributes the Lower Vogelsberg (*LVB*) holds an intermediate feature space between the Vogelsberg (*VB*) and the Wetterau (*WNE*).

Generally, all soilscapes can be described and distinguished by interpreting the terrain features. Thus they should be treated separately in digital soil sensing and mapping approaches.

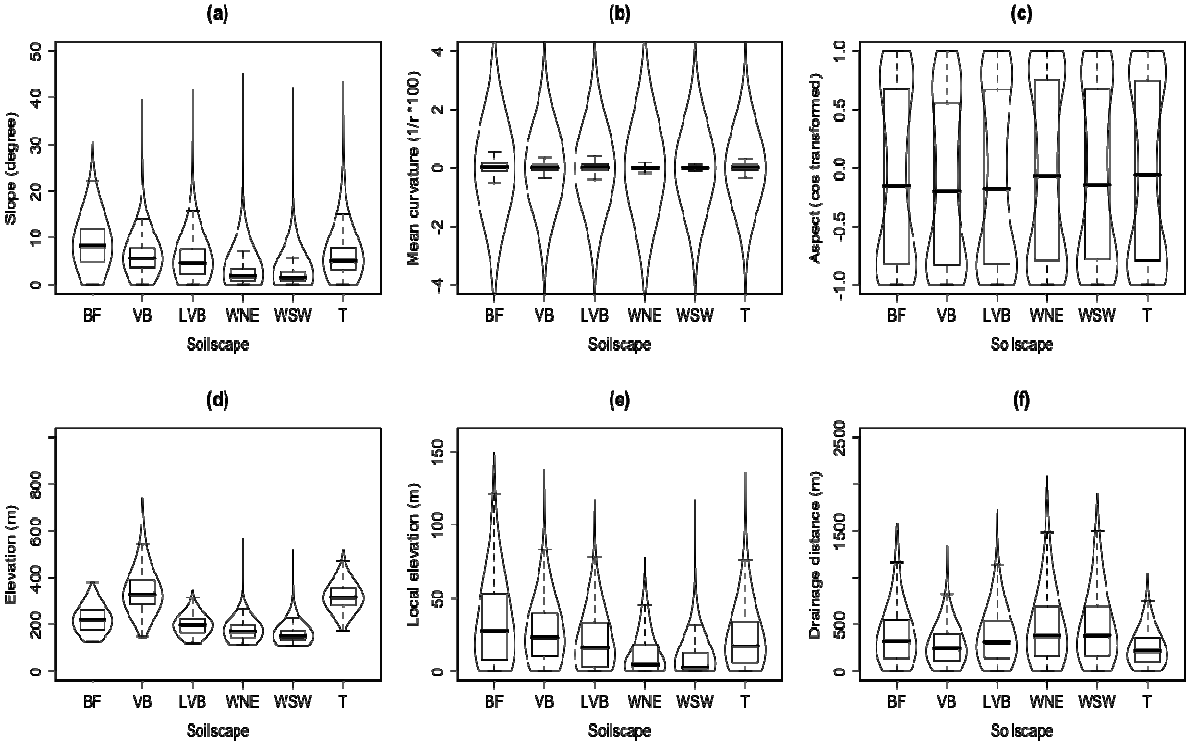


Fig. 5. Feature space analysis (box plot and kernel density plot) for relevant terrain attributes (slope (a), local elevation (b), aspect (c), drainage distance (d), mean curvature (e)) for the soilscapes: Forest of Büdingen (*FB*), Vogelsberg (*VB*), Lower Vogelsberg (*LVB*), Wetterau (*WSW*), Wetterau (*WNE*) and Taunus (*T*).

4.3 Comparison of soilscapes and landscape units

The major units of the map of landscape units in the Nidda catchment are: the Hohe Vogelsberg (HVB), the Untere Vogelsberg (UVB), the Vordere Vogelsberg (VVB), the Wetterau (W), the Hohe Taunus (HT), the östliche Hintertaunus (ET), the Büdinger Wald (BF), and the Büdinger-Meerholzer Hügelland (BML) (Fig. 3.1).

Fig. 6 shows the overlay of the landscape units with the soil map and the geological map in comparison to the soilscapes based on the original legends provided by the HLUG for visual

interpretation. Important results of the statistical analysis are provided in the following.

The Vorderer Vogelsberg (VVB) cannot be correlated to the soilscapes derived, as there is no clear differentiation in terms of the distribution of the dominating soils. Like the Untere Vogelsberg (UVB) it is also dominated by Cambisols (25 %) and Stagnic Luvisols (15 %). Thus, in terms of sampling both landscape units overlap and would therefore produce redundant and uncertain information.

Geologically the Büdingen-Meerholzer Hügelland (BML) (cf. Chapter 4) could be differentiated from the Büdinger Wald (BF) and the Wetterau (W). Yet the high amount of Loess as parent material for soil formation produced a Stagnic-Luvisol association (22 %) as found in the Untere Vogelsberg (UVB) and the Northeast Wetterau (*WNE*).

The Wetterau (W) as mapped in the map of landscape units is dominated by Humic-Luvisols (13 %) and Luvisol-Tschernosems (8 %). This corresponds almost exactly to the soilscape Southwest Wetterau (*WSW*). Yet, the latter is smaller and more homogenous as two dominating soils have a higher percentage (17 % and 11 % compared to 13 % and 9 %). Thus, the differentiation on terms of the distribution of soils provided by the map of landscape units is not as precise as achieved by the proposed approach.

The Hintertaunus (ET / *T*) and the Forest of Büdingen (BF / *FB*) can be found in both the map of landscape units and the map of soilscapes and do overlap in large extents.

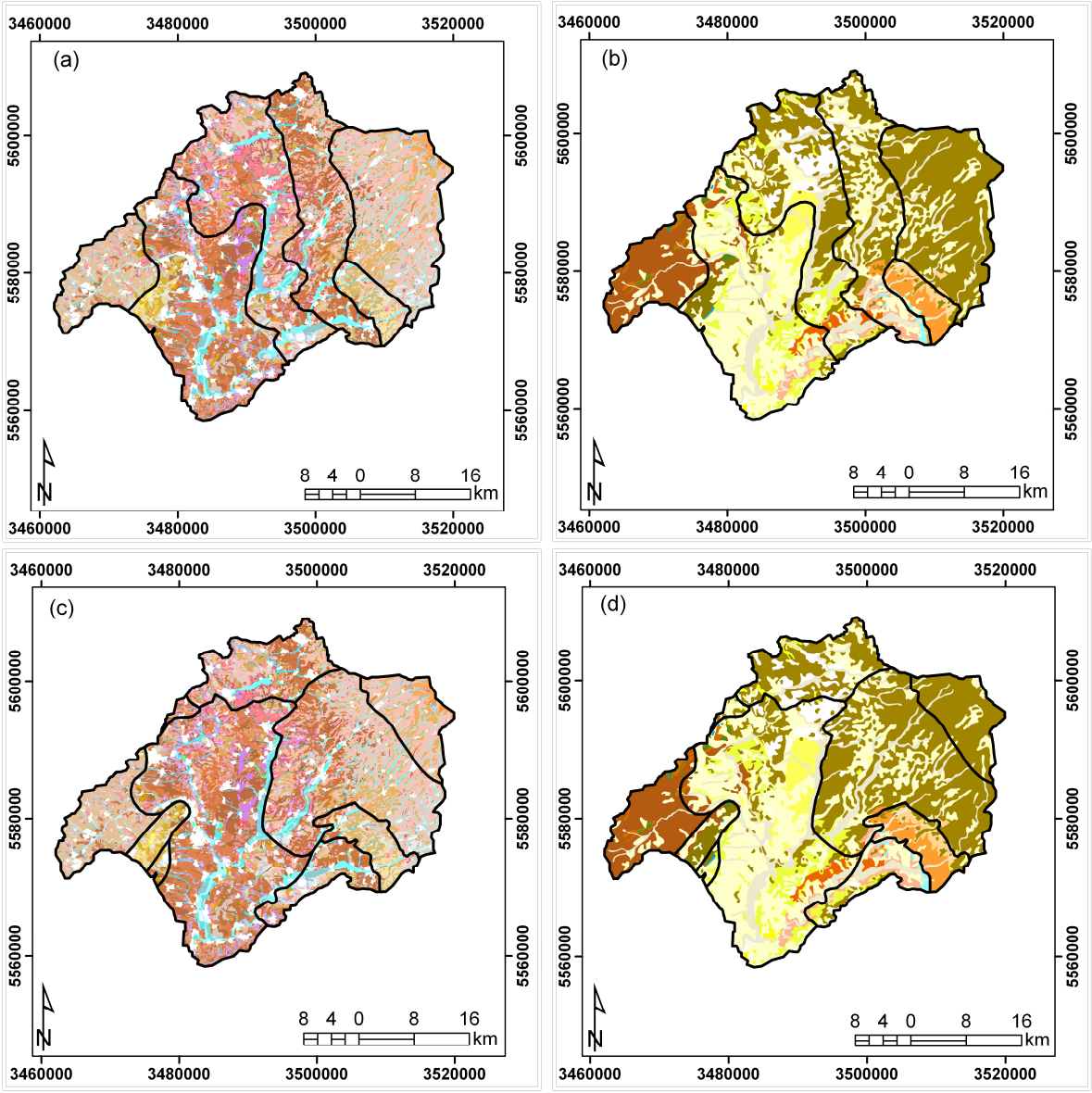


Fig. 6. Comparison between the soilscapes (a, b) and the geographical classification of natural landscapes (c, d) with the geological map 1: 300.000 (b, d) and the soil map 1:50.000 (a, c).

The analysis shows that compared to the map of soilscapes the landscape units does not differentiate the distribution of soils as well, which is due to the different concepts and the fields of application.

Some small units provided by the map of landscape units cannot be found in the map of soilscapes. This is a result of the statistical approach and an important limitation which has to be addressed in further segmentation approaches. A simple solution might be to extend the area, as the small landscape units in the Nidda catchment are offshoots of larger areas.

4.4 Comparisons to the units of the geological map

The overlay of the soilscapes with the geological map reveals a strong visual correlation between the soilscapes on the one hand and the distribution of parent material on the other (Fig. 6). This serves as a validation scheme of the plausibility of the approach proposed. Following the discussion above, additional landscape units might have been derived in terms of differences in parent material. This is for example the case for the Büdingen-Meerholzer Hügelland (BML) in the southeast as accounted for in the map of landscape units (Fig. 6), and also for the so-called Vorder-Taunus. Yet, in terms of size and spatial distribution there might have been additional units which could have been differentiated on the same aggregation level. One example is the Heldenberger Wetterau (located in the Wetterau (W) at the boundary to the Büdingen-Meerholzer Hügelland (BML)) as shown in the next hierarchical level of the map of landscape units, which is characterized by sandstones in large parts. Thus, - including the analysis on Chapter 3.1 – both the map of soilscapes as well as the map of landscape units might be differentiated in more detail. Yet, in terms of the basic idea behind the study and the strong correlation to the distribution of the geological units the aggregation level of the map of soilscapes seems to be appropriate based on the boundaries of the Nidda catchment.

The comparison of the spatial distribution of substrates for each soilscape and each landscape unit reveals surprisingly high similarities between:

- the Forest of Büdingen (*FB*) and the Büdinger Wald (BF) with more than 60 % of sandstones,
- the Taunus (*T*) and the Östliche Hintertaunus (ET) with about 75 % of shale and sandstone,
- the Vogelsberg (*VB*) and the Hohe Vogelsberg (HVB) with 75 % to 80 % of Basanite and Alkaline Basalt,
- the Northeast Wetterau (*WNE*) and the Vordere Vogelsberg (VVB) with 30 and 47 % of Basanite and about 18 % of Loess,
- as well as the Southwest Wetterau (*WSW*) and the Wetterau (W) with 46 to 50 % of Loess and 14 % of fluvial sediments.

Within the map of landscape units the Untere Vogelsberg (UVB) and the Vordere Vogelberg (VVB) are very similar with 57 % and 47 % of Basanite and Alkaline Basalt and 23 % and 18 % of Loess influenced substrates.

The Southwest Wetterau (*WSW*) soilcape which strongly correlates with the Wetterau (*W*) of the map of landscape units contains 5 % more Loess substrates than the latter. It is therefore separated better from the surrounding units.

Thus, even on the basis of the Geological map the delineation of the soilscapes seems more plausible than the landscape units provided in the map of landscape units.

4.5 Technical discussion

The approach to segment a soil map into soilscapes is mainly based on two parameters which have to be fitted: the size of the moving window and the number of classes or segments. Depending on the landscape situation the number of classes is the major component. In this study it is based on the analysis of shape parameters and the aim to derive meaningful segments as small as possible. Thus, it can be seen as a flexible approach in landscape segmentation.

A crucial point is the fragmentation of the soilscapes. For feasibility fragmentation should be as low as possible. However, abrupt changes in soil distribution might be smoothed out due to low fragmentations. Thus, results should be interpreted in terms of fragmentation.

Additionally to fragmentation, further research in landscape segmentation based on existing soil maps should also be tested accounting for taxonomical distances between the soil classes for better delineations that might be shifted due to a soil class and moving window based approach.

As described above, small soilscapes, which are important in terms of soil taxonomy but underrepresented in spatial extent, are not accounted for due to the statistical approach based on frequency distributions. Hence, regions outside the current study area should also be included in the segmentation approach to avoid this problem.

5 Conclusions

In this paper we propose an approach to segment landscapes into soilscapes based on existing soil maps. The aim was to derive homogeneous, non-fragmented units with well-defined soil associations. The approach was developed as the first step of a stacked sampling scheme as a basis for digital soil sensing and mapping within the Collaborative Research Centre 299 of the German Research Foundation.

Based on a spatial moving window frequency analysis in combination with a k-means cluster analysis we derived six representative soilscapes, which, in terms of the geological and topographical homogeneity seems to be more appropriate than the delineation of landscape untis as found in the map of landscape units. This mainly results from focussing on pedological aspects instead of (additional) parameters such as species numbers and distributions. Hence, the proposed approach is superior to characterize soilscapes and thus suitable as a basis for subsequent separated sensing and mapping approaches.

The method might also be used to characterize landscapes for various purposes in landscape-scale soil and environmental research (Pennock and Veldcamp, 2006) from hydrological mapping (Sonneveld, et al., 2006) to pedogenetic and soil-landscape evolutionary models (Minasny and McBratney, 2006).

Methodologically, taxonomical differences between soil classes should be considered in future studies to allow for optimized delineations.

Acknowledgements

Funding for this research was provided by the Collaborative Research Centre 299 of the German Research Foundation. We would like to thank the Hessian State Office for Environment and Geology (HLUG) for providing data.

References

Behrens, T., Schneider, O., Lösel, G., Scholten, T., Hennings, V., Felix-Henningsen, P., Hartwich, R. (2008a): Analysis on pedodiversity and spatial subset representativity – the German soil map 1:1.000.000. J. Plant Nutr. and Soil Sci., (in press).

Behrens, T., Schmidt, K., Gerber, R., Albrecht, C., Felix-Henningsen, P., Scholten, T. (2008b): Concepts for generating shortest representative transects - sampling approaches for linear operated proximal soil sensing techniques. J. Geogr. Inf. Science, submitted.

Behrens, T., Schmidt, K., Zhu, A-X, Scholten, T. (2008c): Multi-scale digital terrain analysis and feature selection in digital soil class mapping. Geoderma, submitted.

Buol, S.W., Hole, F.D., McCracken, R.J. (1973): Soil genesis and classification, Univ. Press, Ames. IA.

Butler, B.E. (1959): Periodic phenomena in landscapes as a basis for soil studies. Commonwealth scientific and industrial research organization, Australia soil publication 14, 20 pp.

Dikau, R. (1992): Computergestützte Geomorphologie. - Habilschr. Fak. Geowiss., Univ. Heidelberg: 303p.

Everitt, B.S. (1980): Cluster analysis (2nd ed.). Wiley, New York.

Friedrich, K. (1996): Digitale Reliefgliederungsverfahren zur Ableitung bodenkundlich relevanter Flächeneinheiten. Frankfurter Geowissenschaftliche Arbeiten, Serie D - Physische Geographie, Band 21.

Gerber R., Salat, C., Junge, A., Felix-Henningsen, P. (2007): GPR-based detection of Pleistocene

periglacial slope deposits at a shallow-depth test site. Geoderma 139, 346-356.

Gerrard, A.J. (1980): Soils and landforms – an integration of Geomorphology and Pedology. George Allen and Unwin, London. 216p.

Hessian State Office for Environment and Geology (2008): Umweltatlas Hessen. http://atlas.umwelt.hessen.de/atlas.

Hole, F.D. (1978): An approach to landscape analysis with emphasis on soils. Geoderma 21, 1-23.

Horn, B. K. P. (1981): Hill Shading and the Reflectance Map, Proceedings of the IEEE, 69(1), 14-47.

Jenny, H. (1941): Factors of soil formation. McGraw-Hill: New York, 281p.

Kalis, A.J., Merkt, J., Wunderlich, J. (2003): Environmental changes during the Holocene climatic optimum in central Europe – human impact and natural causes. Quarternary Science Reviews 22 (1), 33-79.

Klausing, O. (1988): Die Naturräume Hessens mit einer Karte der naturräumlichen Gliederung 1:200000. Schriftenreihe der Hessischen Landesanstalt für Umwelt 67, Wiesbaden. 43p.

Kundert, K. (1988): Untersuchungen zur automatischen Klassifikation von räumlichen Einheiten. – Geoprocessing, 7, Zürich. 113p.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, Fifth Berkeley Symposium on Math Statistics and Probability, 281-297.

McBratney, A.B., Hart, G.A., McGarry, D. (1991): The use of region partitioning to improve the representation of geostatistically mapped soil attributes. Journal of Soil Science 42, 513-532.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B. (2003): On digital soil mapping. Geoderma 117. 3-52.

McSweeney, K., Slater, B.K., Hammer, R.D., Bell, J.C., Gessler, P.E., Petersen, G.W., (1994): Towards a new framework for modeling the soil-landscape continuum. in Amundson, R., Harden, J., Singer, M.: Factors of soil formation: A fifthieth anniversary retrospective. SSSA Special Publication 33, 127-145.

Meynen, E., Schmithüsen, J. (1953): Handbuch der naturräumlichen Gliederung Deutschlands. Gemeinschaftsveröffentlichung des Instituts für Landeskunde und des Deutschen Instituts für Länderkunde. Bad Godesberg.

Milne, G. (1935): Some suggested units of classification and mapping, particularly for East African soils. Soil Res. 3, 183-198.

Möller, M., Volk, M., Friedrich, K., Lymburner, L. (2008): Placing soil genesis and transport processes into a landscape context: A multi-scale terrain analysis approach. J. Plant Nutr. and Soil Sci. 171, 419-430.

Pennock, D.J., Veldkamp, A. (2006): Advances in landscape-scale soil research. Geoderma 133, 1-5.

Pullan, R.A. (1969): The soil resources of West Africa. in Thomas, M.F., Whittington, G.W. (Editors), Environment and Land Use in Africa. Methuen, London, 147-191.

Ruhe, R.V. (1956): Geomorphic surfaces and the nature of soils. Soil Science, 82/6, 441-455.

Schmidt, K., Behrens, T., Scholten, T. (2008): Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma, 146, pp. 138-146.

Schmidt, R. (1975): Grundlagen der Mittelmaßstäbigen Landwirtschaftlichen Standort-kartierung. In: Arch. Acker- u. Pflanzenbau u. Bodenkd., Berlin 19, 8, S.533-543.

Sonneveld, M.P.W., Schoorl, J.M., Veldkamp, S. (2006): Mapping hydrological pathways of phosphorus transfer in apparently homogeneous landscapes using a high-resolution DEM. Geoderma 133, 32-42.

Stewart, G.A., Perry, R. A. (1953): Survey of Bowen-Townsville region (1950), CSIRO Australia, Land Research Series, 2,87pp.

Tarboton, D. G. (1997): A New Method for the Determination of Flow Directions and Contributing Areas in Grid Digital Elevation Models. Water Resources Research, 33(2): 309-319.

Webster, R., Beckett, P.H.T., (1968): Quality and usefulness of soil maps. Nature 219, 680-682.

Werban, U., Behrens, T., Cassiani, G., Dietrich, P. (2008): iSOIL – An EU-project for Integration of Geophysics, Digital Soil Mapping and Soil Science. In McBratney, A., Viscarra-Rossel, R. (Ed.): Global Workshop on High Resolution Digital Soil Sensing and Mapping, Sydney (in press).

Zevenbergen, L.W., Thorne, C.R. (1987): Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12: 47-56.

Manuscript 2

Concepts for generating shortest representative transects – sampling approaches for linear operated proximal soil sensors

Journal of Geographical Information Science, submitted in March 2009

T. Behrens[1], K. Schmidt[1], R. Gerber[2], C. Albrecht[2], P. Felix-Henningsen[2] and T. Scholten[1]

[1] Institute of Geography, Chair of Physical Geography, Eberhard Karls University, Rümelinstraße 19-23, D-72074, Tübingen, Germany

[2] Institute of Soil Science and Soil Conservation, Justus Liebig University, Heinrich-Buff-Ring 26-32, D-35392, Giessen, Germany

Abstract

This paper introduces two concepts for generating representative transects as a basis for optimized linear operated proximal soil sensing surveys in large areas. The general approach is based on nominal spatial datasets in terms of stratified sampling. The aim is to provide transects covering all relevant feature combinations within an investigation area over the shortest possible distance, in order to obtain typical, valid and transferable datasets for spatial predictions in digital soil mapping. The first concept focuses on finding a single shortest transect based on randomized searches whereas the second concept generates multiple transects and follows the idea of a more holistic information retrieval such that all relevant interclass transitions of the spatial stratification could be recorded – again over the shortest possible distance(s). The crucial step in computing transects is the location of the transect nodes. We compare and discuss an approach based on the centre point of each class-area in a map that is furthest from the boundaries and an approach where a centreline of points is generated for each class-area.

The methods are applied and compared within representative soil map patches as stratification information of six soilscapes in the Nidda catchment (Hesse, Germany). The random component to delineate single transects is additionally tested on 9 artificial random datasets.

The results show that the multiple transect approach returns longer transects, but covers the feature space in terms of the frequency distribution of underlying terrain attributes more comprehensive as it is closer related to the catena concept when applied on an existing soil map. In terms of computation time, transect length, and feature space description, a combination of the multiple transects and the centreline of points approaches must be recommended. However, there is a trade-off between transect length on the one hand and computation time and feature

space description on the other hand which might lead to choose the single transects concept based on the centreline approach. Thus, the decision whether to use a single transect or multiple transects depends on comparing cost effectiveness and obtainable information.

Representative shortest transect sampling can be useful for data collection in digital soil mapping frameworks in order to achieve valid prediction results when linear operated proximal soil sensing techniques are used. Furthermore, the multiple transects approach can be used to investigate important soil-association transition zones in a landscape and can serve as a tool for boundary quality estimations of existing soil maps.

1 Introduction

Based on the economic and ecologic pressure to estimate and handle the impacts of global climate change, the demand for high resolution soil property data for large areas is strong and growing. In contrast, efficient sampling and mapping at high resolutions < 20m for areas comprising more than 100 km^2 is challenging. One of the most promising approaches is combining proximal soil sensing and digital soil mapping techniques (McBratney et al., 2000, 2003; Behrens and Scholten, 2006; Viscarra-Rossel et al., 2007). Opposed to the field scale surveys for precision agriculture where high density soil sensing is easily possible the application of proximal soil sensing techniques is ambitious for larger areas especially when operated along transects in larger soilscapes and across different landuse classes. Thus, the first step for a fast and cost effective creation of high resolution soil property data for large areas is to generate optimized transects, which should be as short as possible and representative for the investigation area.

In ecology line transect sampling (Burnham et al., 1980; Jensen, 1996; Manly, 2002) is a well-known method to acquire spatially distributed and correlated information (Hedley and Buckland, 2004) where most approaches focus on estimating the density of animal or plant populations in a study area (Stoyan, 1982; Drummer and McDonald, 1987). In soil science the catena concept is a well-known approach to derive qualitative and quantitative models (Hoosbeek and Bryant, 1992; Sommer and Schlichting, 1997; McBratney et al., 2003). Transects have been used for soil surveys especially in areas with poor access (Acres et al., 1993). For reconnaissance soil surveys in tropical forests Acres et al. (1993) developed a sampling scheme to locate transects with starting points randomly located along motorable roads within different land units. Another approach for transect soil sampling was described by Hengl et al. (2003) who constructed a transect in the direction of the azimuth of the hightest anisotropy in the feature space. Discussions on techniques to generate representative transects as a basis for linear operated proximal soil sensing techniques like ground penetrating radar or electro-

magnetic surveys has been limited.

In this methodological paper we introduce two different approaches to generate shortest representative transects, which can be applied on nominal datasets of all scales and resolutions. In this respect, representativity is defined as covering all relevant information within a study area with the aim to obtain valid and transferable measurements for subsequent digital soil mapping (DSM) approaches (cf. Behrens et al. 2008a). We apply the approach on existing soil class maps to generate high resolution soil property maps.

2 Preliminaries

2.1 Transect definition and sampling

In contrast to k-th order random toposequences, a newly introduced approach by Odgers et al. (2008) as a sampling strategy for DSM, which are monotonically downhill transects calculated directly on the basis of a digital elevation model (DEM) and thus are related to the classical catena concept (Milne, 1935; Sommer and Schlichting, 1997), our approaches are based on nominal datasets and thus belong to the group of stratified sampling schemes (McKenzie and Ryan, 1999; de Zorzi et al., 2005). Hence, the transects derived are neither directly bound to a topographical gradient, as for catenae and toposequences, nor restricted to a process based linkage of soil units and uniform parent material as implied for a catena (Milne, 1935; Bushnell, 1942; Odgers et al., 2008).

In terms of a stratified sampling scheme, existing soil maps as well as combined maps of environmental co-variates like geology, landuse and relief might be used for the purpose of optimised soil surveys and soil sensing approaches (McKenzie and Austin 1993; Gessler, 1995; McKenzie and Ryan, 1999; Gerber et al., 2007). In this study we use existing soil class maps for stratification aiming to collect high resolution spatial soil attribute data.

2.2 Basic concepts for transect generation

Technically, we differentiate a single transect concept and a multiple transects concept. The single transect concatenates all classes occurring in a study area into one single transect where each class is represented once. In contrast, the multiple transects approach returns all shortest transects found between adjacent map classes to cover all relevant continuous transitions zones, and is therefore a more holistic approach in terms of landscape characterisation and closer related to the catena (Milne, 1935) and the toposequences (Gobin et al., 2000; Odgers et al., 2008) concepts.

In contrast to the multiple transects approach which is computationally easy, the task of

generating a single concatenated representative shortest transect is more complex. In principle, three approaches seem possible:

The first is a greedy search of all possible transects in a dataset, which results in *l* transects:

$$l = \prod_{c=1}^{n} f_c * \frac{n!}{2} \quad (1)$$

where f_c is the frequency of each class *c* and *n* the number of classes in a dataset.

A second approach possible is based on graph-theory (e.g. Djikstra, 1959). Compared to the travelling salesman problem, shortest path searches or minimum spanning trees (e.g. Kruskal, 1956; Gutin and Punnen, 2006), finding the shortest representative transect is a more generalised task as not all available points are included in the final transect, classes can occur more than once, and neither the start nor end point are known a priori.

Third, as greedy search approaches over all possible transects are not applicable due to an enormous search space even for relatively small datasets and developing approaches based on graph theory are out of the scope of this paper, this study introduces algorithms based on randomised nearest neighbour searches to approximate the shortest single representative transect.

2.3 Embedding transects in DSM approaches

In terms of DSM the transect approach can be used for different kind of purposes. The major objective behind this study is to digitally map soil property data based on calibrated sensor data. Therefore, reference measurements are needed and the sensor data has to be adjusted for that model. For ground penetrating radar it is the conversion of the time function to a depth function. The data sensed along the transect will then be resampled on the basis of the desired resolution of the DSM approach. Based on this dense set of "sample points" and environmental covariates like terrain attributes (Behrens et al., 2008b) a map of soil depth can be digitally mapped using machine learning approaches (McBratney et al. 2003; Grimm et al., 2008). Thus, if for example a coarse scale soil map is used to generate the transect(s) it can, but does not need to be included in the subsequent prediction approach. In this case it serves as a proxy for the distribution of soil properties. On the other hand, one might argue to use the same attributes for transect generation and prediction for consistency. However, we think that this is for sure an option, yet an existing soil class map at an appropriate scale should give the best basis for transect generation, as soil is the result of all state factors (Jenny, 1941). Thus, a combination of a soil map and some additional environmental covariate might bias the model towards this data.

Additionally, it will result in more units and thus longer transects. Furthermore, including a coarser scale soil (or stratified) maps used for generating the transects in the prediction approach might be contradictory because of the desired higher resolution if the boundaries of the soil map have not been adjusted before (Behrens et al. 2008c). However, in the case a soil map is not available, any other reasonable stratification scheme might be used.

Another purpose of the transect approach might be to analyse and/or refine the information content of the soil map used for transect generation including evaluations of the soil boundary delineation quality. In these cases the soil map plays an important role in transect generation and modelling.

The idea behind this study is based on the first objective, the production of soil property maps.

3 Materials and methods

3.1 Transect generation

3.1.1 The node problem

Due to the fact that transects need start points, end points, and nodes, all transect searches are based on point datasets created on the basis of class-area maps from nominal spatial datasets. We propose two approaches to derive the centre point(s) of each class-area (subsequently forming the nodes of each transect).

First, instead of the centre of mass as implemented in many GIS products (e.g. ESRI, 2004), we introduce an approach based on Euclidean allocation to derive that point within a mapping unit, which is located furthest away from the class area boundary, i.e. the centre point of the largest circle which can be placed inside a unit (Figure 1). This ensures that the class-areas are not covered at their uncertain or fuzzy boundaries only, which otherwise might result in non-representative feature values.

Second, based on the same concept, we derived a centreline of points. Therefore, we applied a morphological thinning approach combined with a buffer to avoid points that are too close to the boundary (Figure 1). The centrelines were evenly divided into points based on a user-defined distance.

The technical advantages of the single point approach are the limited set of points and the maximized distance to the boundary. For the centre line approach it is the possibility to compute shorter transects due to more points available, especially for elongated areas.

To avoid single transects that are too long due to very small classes not relevant in terms of representativity, we propose to include only those units that fall within a cumulative frequency

distribution of 96 %. For the multiple transects approach we computed transects only for those neighbouring classes that share a boundary longer than the average minus two times the standard deviation of all boundaries between two soil classes. These thresholds must be regarded as rules of thumb as they strongly depend on the fragmentation of the classes and the amount of tolerable fieldwork.

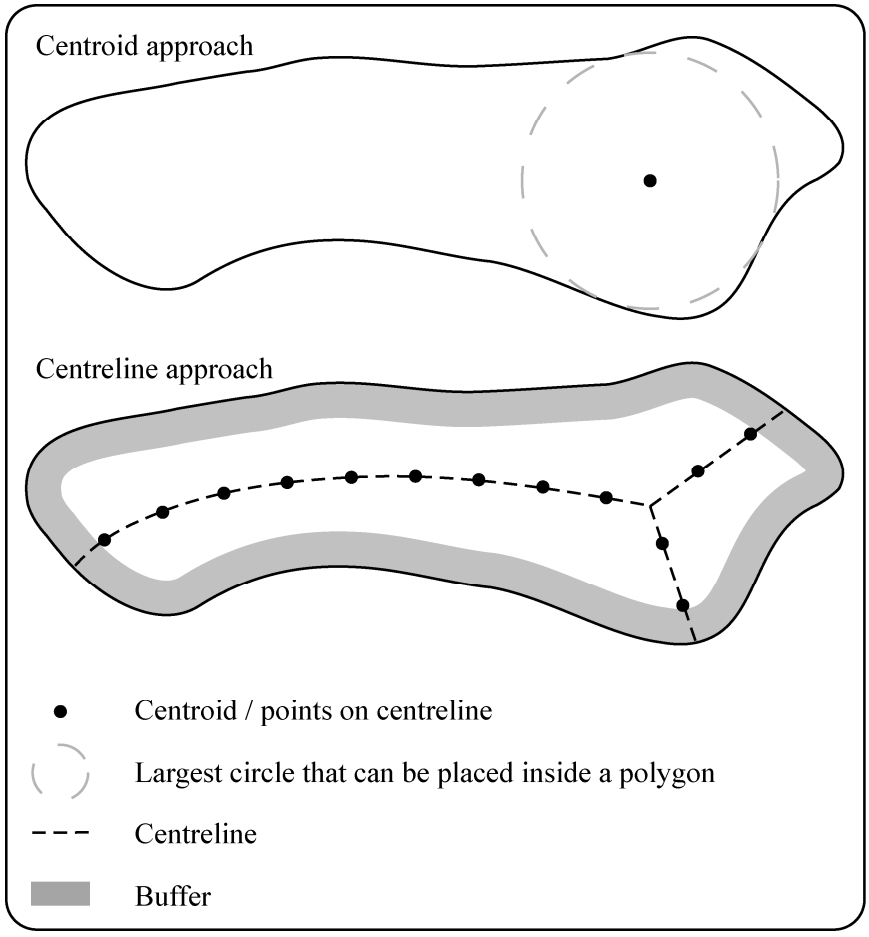


Fig. 1. Illustration of the approaches to derive the centroid and the centreline of points for a polygon.

3.1.2 Single Transects

The main constraints for forming a single representative transect are:

- each class must be included,
- each class must be included only once,
- the transect should be the shortest possible in the dataset.

To reduce computation time compared to greedy-search approaches over the entire transect space, the number of possible transects per starting point has to be reduced. The computationally fastest way is to calculate one transect per point only. Starting at one point, the

nearest neighbour of another class not yet included in the transect is searched. Then the nearest neighbour of the second point is searched. This procedure is repeated until all classes are included in the transect once (NN-search). The pseudo-code is shown in Figure 2.

```
// NN - nearest neighbour search
for each point in a dataset
  while (not all classes are included in the transect)
        find the nearest neighbour of another class not yet included
        sum the distances for each transect
  end
  save the transect and its length
end
sort the list of transects based on their length
return the shortest transect
```

Fig. 2. Pseudo code for finding the shortest transect based on nearest neighbour search.

The major shortcoming of the NN-search approach is that it does not necessarily return the absolute shortest transect, but a more or less vague approximation, as a transect segment does not need to start or end at it's nearest neighbour, but, for example, the next – the second-nearest – neighbour. Thus, we extended the simple nearest neighbour approach with a random component where "nearest" is randomly altered between the 1st and the nth nearest neighbour (RNN-search; Figure 3).

```
// RRN - random nearest neighbour search
for each iteration in 1..i
  randomly select a point
  while (not all classes are included in the transect)
    randomly select one of the n nearest neighbours of another class
    sum the distances and the coordinates for each transect
  end
  save the transect and its length
end
sort the list of transects based on their length
return the shortest transect
```

Fig. 3. Pseudo code for finding the shortest transect based on random nearest neighbour search.

As there is a random component included in the RNN-search method, each starting point in a dataset needs to be tested more than once to find the shortest transect. Additionally, testing each point as a staring point multiple times might result in a much extended computation time. Thus,

we also use randomly selected points as starting points. Without knowledge about the dataset, the shortest transect computed must therefore be regarded as an approximation. Hence, the more iterations tested the higher the probability to find the shortest transect.

3.1.3 Multiple transects

The general technical aims of computing a single transect are identical for the multiple transects approach in terms of representativity and shortness. The main difference is that in the multiple transects case *n* single transects - each consisting of one segment only - are calculated whereas in the single transect approach one multiple segment transect is generated.

In terms of landscape description, transition zones play a crucial role, especially if the sampled data should be used in DSM approaches (cf. Hengl et al., 2003; Odgers et al., 2008). Thus, all relevant soil-associations and their transitions should be covered by independent transects.

Technically, both the nominal class-area dataset as well as their centre points need to be analysed in this approach. In a first step, all combinations of adjacent classes are recorded on the basis of a gridded dataset by analysing neighbouring pixels in a moving window approach. Subsequently, the point dataset is scanned to find the shortest possible distance for each relevant class combination.

3.2 Datasets

To evaluate the influence of the number of iterations in the single transect approach, we constructed 9 square artificial random point datasets each consisting of 100 points and 2 to 10 classes respectively. To compare and analyse the single and the multiple transects approaches, we tested them on 6 representative soil map patches (Behrens et al., 2008a), each comprising an area of 3*3 km, at a scale of 1:50,000 and a DEM resolution of 10 meters within 6 diverse soilscapes of the Nidda catchment, Hesse, Germany (Schmidt et al., 2008) (Figure 4).

The six soilscapes were computed on the basis of the official soil map 1:50,000 provided by the Hessian Agency for Environment and Geology to create features spaces in terms of soil associations with a high interclass and a low intraclass variability. The idea behind this approach is to provide homogeneous soilscapes for a GPR-based DSM approach within the Collaborative Research Centre (SFB) 299 of the German Research foundation (DFG) (Schmidt et al., 2008). The soilscapes are:

- the *Forest of Büdingen* characterized by bunter (BF),
- the *Taunus*, as part of the Rhenish Massif with predominantly slates (T),
- the *Upper Vogelsberg*, which forms the core of the largest basalt formation in Central Europe (VU),

- the *Lower Vogelsberg*, a landscape influenced by basalt covered with Pleistocene periglacial slope deposits and loess depositions (VL)
- and two different loess soilscapes within the *Wetterau* (WN, WS)

The representative patches within these soilscapes were computed on the basis of a χ^2-moving window similarity test aiming to minimise the investigation area while preserving information content (Behrens et al., 2008a). Thus, each patch, a transect is calculated for, comprises a unique feature space.

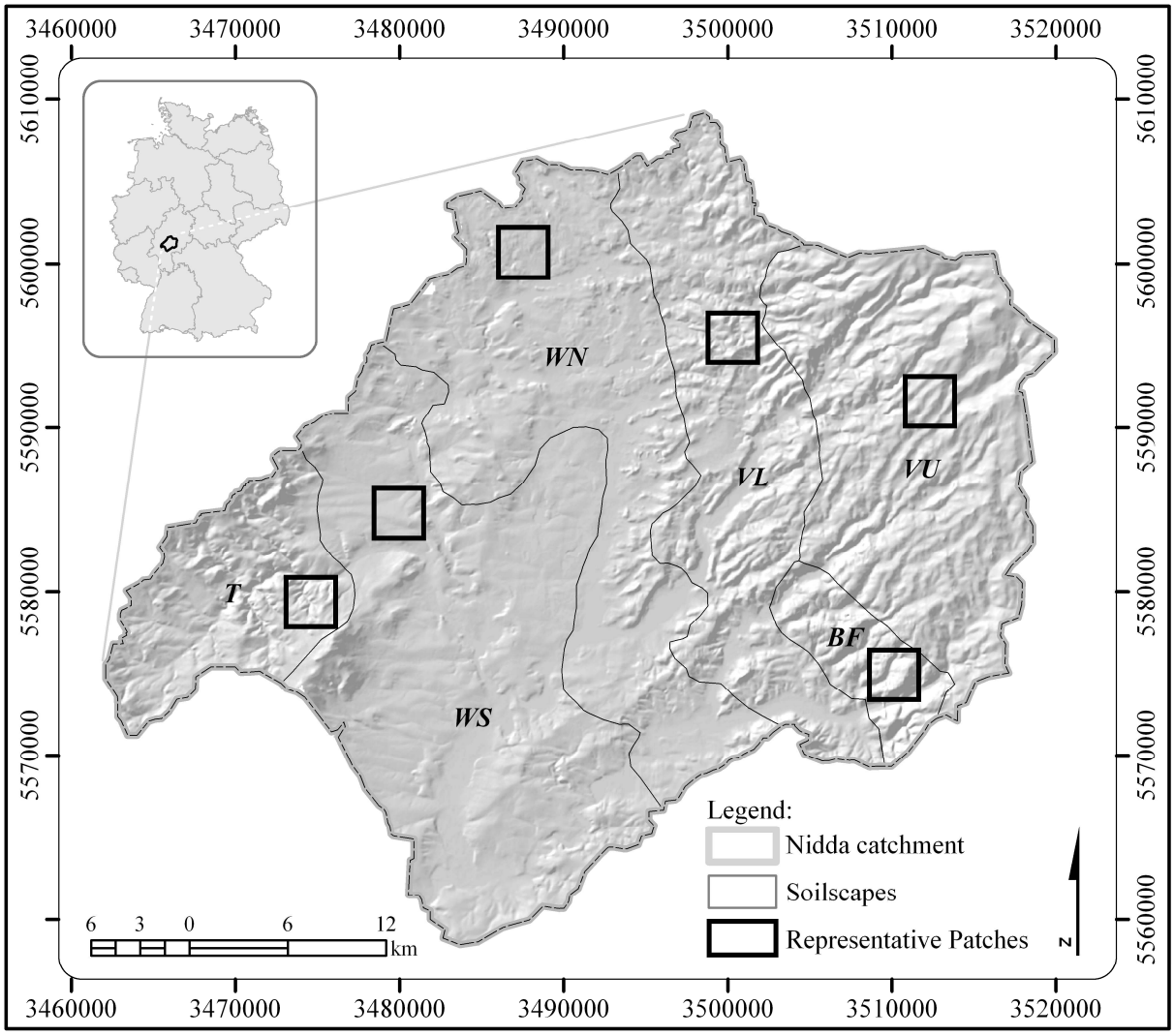


Fig. 4. Soilscapes and representative patches (draped over a hillshade) of the Nidda catchment used to compare the shortest representative transect approaches. *BF* = Forest of Büdingen; *T* = Taunus; *VU* = Upper Vogelsberg; *VL* = Lower Vogelsberg; *WN* = Wetterau (north-east); *WS* = Wetterau (south-west).

3.3 Computation

The computation and comparison focuses on all combinations of single and multiple transects and the centroid as well as the centreline approaches.

3.3.1 Analysis Settings

To analyse the single transect approach we tested 100, 1,000, 10,000, 100,000, and 1,000,000, iterations on the basis of 2, 3, 4, 5, 6, 7, 8, 9, and 10 random nearest neighbours on the artificial datasets. The comparisons of transect length and computation times are based on the six different soilscape patches described in chapter 7. For these real world datasets 1,000,000 iterations were used to construct the single transects.

For deriving the centreline of points we used a point spacing of 50 meters and a buffer of 20 m inside each class-area.

3.3.2 Feature space analysis

To compare the different transect generation methods we analysed the feature spaces (Lillesand and Kiefer, 2000) of terrain attributes covered by the transects and compared these feature spaces with the ones for the entire patches. Hence, the more similar the feature space covered by a transect compared to the underlying patch is, the more reasonable the approach.

The feature space was analysed in terms of the mean, the standard deviation, and the range of slope, cosine transformed aspect, and relative hillslope position (cf. Behrens et al., 2008b).

4 Results and discussion

4.1 Artificial datasets

The single transect approach needs two parameters to be specified: the number of iterations and the number of neighbours included in the random search approach which both have an influence on computation time. If the number of neighbours is large and/or if the number of iterations is large, computation time rises. On the other hand, if there are only few iterations and/or the number of neighbours is large, the probability to find the shortest possible transect decreases.

Figure 5 shows the results for the number of neighbours on the random datasets. In most cases, the first nearest neighbour approach returned the shortest transect for datasets with a small amount of classes. Yet, from 7 classes onward at least 2 random nearest neighbours were needed to return the shortest possible transect (Table 1).

Table 1 lists the frequencies of the shortest transect detected for the artificial datasets in respect to the number of iterations tested, clearly showing that more iterations are needed with an increasing number of classes. Additionally, it can be seen that about 50 to 100 times as many

iterations are needed if 2 random nearest neighbours are needed, which is the case for all datasets containing seven and more classes.

Even if a full random approach without constraints in the amount of nearest neighbours included would return the shortest transect when enough iterations were applied, using only the second or the third nearest neighbour ensures shorter computation times while returning reasonable approximations in terms of transect length. However, a general function to calculate the optimised values for the *n*-nearest neighbours or the number of iterations needed cannot be given on the basis of the data tested.

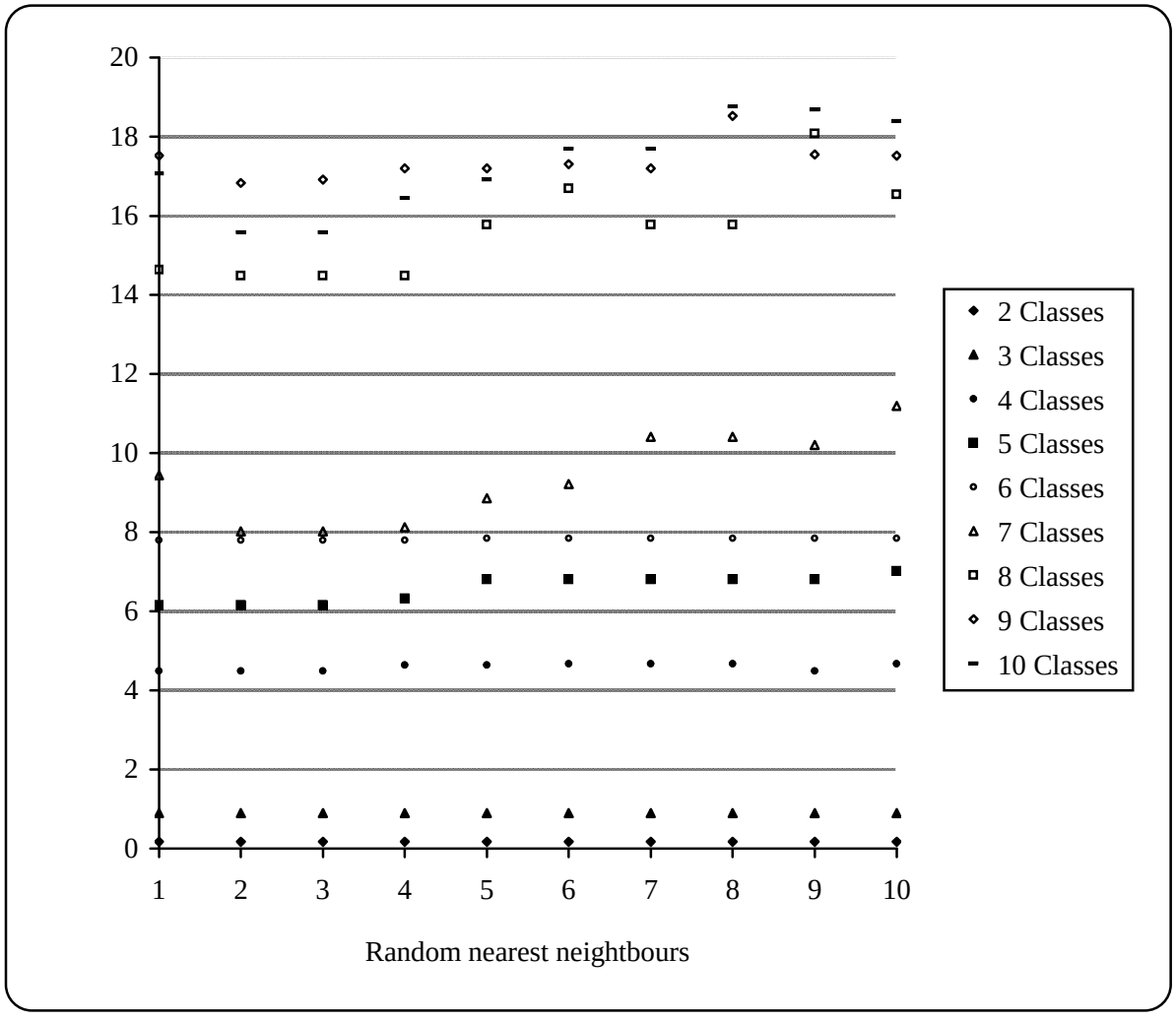


Fig. 5. Transect length (relative units) and random nearest neighbours based on 9 artificial random datasets consisting of 2 to 10 classes.

Further interpretations do not seem appropriate, as all components including the dataset and the algorithm are random. Yet, the general trend and the need to search over n-nearest neighbours is demonstrated.

As there is no random component needed in the multiple transects approach it always returns the shortest transects. Thus, a detailed comparison with the single transects approach is

discussed in Chapter 4.2 on the real world datasets.

Table 1. Amount of shortest transects found for each combination of classes in the dataset and iterations used. Rn = minimum random nearest neighbour to find the shortest transect.

		Classes								
		2	3	4	5	6	7	8	9	10
Iterations	100	10	5	2	2	2	0	0	0	0
	1000	72	30	11	8	11	21	0	0	0
	10000	711	292	93	110	94	3	2	0	0
	100000	7062	3051	961	1019	1055	15	16	0	3
	1000000	70675	30417	9465	10469	11135	170	179	8	8
Rn		1	1	1	1	1	2	2	2	2

4.2 Real datasets

4.2.1 Transect length and computation time

The primary result to be discussed is the length of the transects computed. Table 2 shows the comparison for the six soil map patches. In average, the length of the multiple transects (in total) is twice as long as the length of the single transects. Thus, with soil sampling or soil sensing being the most expensive part of a survey (Webster and Oliver, 1990), the single transect approach must be recommended. The overall average of 3.300 m for the centreline approaches and 4766 m for all approaches seems to correspond to the patch size of 3 * 3 km and its diagonal length of 4200 m respectively. Further studies on different patch sizes and map scales have to prove if this could be used as an approximation of the resulting transect length.

It can be seen that the average length of the transects varies strongly between the approaches. The transects computed using the multiple transects centroid approach are about 3.5 times longer than the ones computed using the single transect centreline approach which generally returns the shortest transects. Compared to the centroid approaches the centreline approaches were 21 % shorter in case of the single transect and 53 % shorter in case of the multiple transects. The multiple transects centreline approach returned transects that are 26 % longer compared to the single transect centreline approach. Additionally, the actual distance travelled will be much greater for the multiple transects approach even if a sensor is not being deployed between the transects sections. Hence, in terms of feasibility the single transects approach based on the centreline method must be recommended.

Generally, the transects returned are relatively long for proximal soil sensing approaches especially when applied across different landuse classes. Thus, sub-sampling schemes seem appropriate where, for example, sub-transects of 50 m length might be recorded every 100 m. To sample points along the transects, equal distance or equal elevation intervals might be applied (Odgers et al., 2008).

Table 2. Transect length [m] and computation time [s] for the single and the multiple transects approaches. *BF* = Forest of Büdingen; *T* = Taunus; *VU* = Upper Vogelsberg; *VL* = Lower Vogelsberg; *WN* = Wetterau (north-east); *WS* = Wetterau (south-west).

	Single transect				Multiple transects			
	Centroid		Centreline		Centroid		Centreline	
Soilscape	Length [m]	Time [s]	Length [m]	Time [s]	Length [m]	Time [s]	Length [m]	Time [s]
BF	4361	1488	3426	3857	12752	<10	5385	<10
T	1740	1595	1291	2227	10333	<10	5256	<10
VU	1788	523	1683	1186	3657	<10	2437	<10
VL	4839	3787	4585	11417	7589	<10	4300	<10
WN	3675	5480	2605	8670	7524	<10	2913	<10
WS	4260	2834	2638	5950	12332	<10	3003	<10
mean	3444	2618	2705	5551	9031	<10	3882	<10

The computation times for finding the shortest single transect based on 3 random nearest neighbours and 1,000,000 iterations as well as the corresponding times for the multiple transects approach are listed in Table 8.2. The single transect approach needed between 45 min and 90 min in average (on a standard PC, 2.6 GHz, 2GB RAM) to approximate the shortest transect, whereas the multiple transects approaches only needed a few seconds, as no iterative random component is implemented. The larger the node space of a datasets the larger the differences in computation time. Here, the multiple transects approach obviously has a clear advantage.

4.2.2 Feature space analysis

The feature space analysis in terms of landscape representation (Table 3) measured on the basis of a similar frequency distribution of slope [°], aspect (cosine transformed [-1 - 1]), and hillslope position (indicating whether a pixel is located upslope (+10), midslope (0), or downslope (-10)), the multiple transects approach clearly outperformed the single transect approach.

Concerning all four approaches, the multiple transects centroid approach returned the highest

similarity in 23 cases followed by the multiple transects centreline approach (22), the single transects centroid approach (12), and the single transects centreline approach (10). Thus, in terms of covering the feature space of a landscape, the multiple transects approaches must be recommended. Interestingly, this is not directly related to the overall transect(s) length. Comparing the two multiple transects approaches, the centroid approach returns transects that are 53 % longer but only provide a better description of the underlying feature space of 12 %. Regarding the centreline approaches, the difference in distance is only about 26 %, but the difference in the feature space description is 43 %. This furthermore shows that the centreline approach performs well regardless of the fact that the centreline points are generally located closer to the class-area boundaries compared to the centroid approach.

4.2.3 Transect location

Another issue to be discussed is the location of the transect in the map. As shown in Figure 6, the shortest transect between two units - either a transect of the multiple transects approach or a segment of a single transect - might cross other units which is contradictory to the initial idea and often related to the shape of the mapping units. This problem mainly occurs for transects of the single transect approaches and especially for the centroid method due to the fact that only one point is generated per class-area. To provide optimised results in terms of surveying specific soil transition zones, the multiple transects centreline method should be applied.

Table 3. Deviations between the values for the entire patches and the data recorded along the transects in terms of mean, standard deviation, and range of slope, aspect (cosine transformed), and hillslope position for the single transect approach and the multiple transects approach both for the centroid approach and the centreline approach. Bold numbers indicate the smallest deviation. *BF* = Forest of Büdingen; *T* = Taunus; *VU* = Upper Vogelsberg; *VL* = Lower Vogelsberg; *WN* = Wetterau (north-east); *WS* = Wetterau (south-west).

			BF	T	VU	VL	WN	WS
Multiple transects - Centroid	Slope [°]	Range	**4.58**	10.06	7.38	**6.32**	**0.88**	**4.74**
		Mean	0.95	0.62	0.61	**0.15**	-0.66	**-0.19**
		Stdev	0.23	**-0.33**	-0.13	***-0.03***	-0.62	-0.06
	Position [-10 - 10]	Range	**3.13**	3.20	3.51	2.79	3.13	**3.71**
		Mean	0.26	0.55	-0.55	0.11	0.19	0.80
		Stdev	-0.21	-0.09	-0.34	0.14	-0.15	0.16
	Aspect [-1 - 1]	Range	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
		Mean	**0.02**	0.08	**0.21**	**0.03**	-0.14	-0.19
		Stdev	**0.00**	**0.00**	-0.07	**0.02**	**0.06**	**0.01**
Single transect - Centriod	Slope [°]	Range	11.00	10.23	**0.62**	13.51	10.47	15.46
		Mean	-1.17	-1.70	0.32	0.50	-0.77	-0.39
		Stdev	1.36	-1.03	-0.51	0.95	**0.06**	0.18
	Position [-10 - 10]	Range	7.02	7.44	6.73	4.39	3.92	6.21
		Mean	0.41	**0.06**	**0.17**	0.27	**0.15**	**-0.22**
		Stdev	0.52	0.48	**0.16**	0.17	0.18	0.20
	Aspect [-1 - 1]	Range	-0.01	**0.00**	-0.01	**0.00**	-0.10	**0.00**
		Mean	-0.39	-0.03	-0.41	0.13	0.38	**0.06**
		Stdev	0.11	-0.02	**-0.04**	0.03	0.23	-0.03
Multiple transects - Centreline	Slope [°]	Range	6.33	**2.65**	8.23	6.70	9.51	13.41
		Mean	**-0.34**	**-0.06**	0.19	0.23	**0.12**	-0.34
		Stdev	**0.00**	-0.85	**0.12**	**0.03**	0.23	**0.04**
	Position [-10 - 10]	Range	5.14	**2.64**	**3.20**	4.63	**2.99**	4.94
		Mean	1.03	0.48	-0.21	0.24	0.55	1.29
		Stdev	0.10	**0.07**	-0.89	0.14	**0.01**	**-0.05**
	Aspect [-1 - 1]	Range	**0.00**	**0.00**	**0.00**	**0.00**	0.02	**0.00**
		Mean	-0.22	0.03	-0.31	-0.20	**0.11**	0.26
		Stdev	0.08	-0.04	-0.12	0.09	**0.06**	**0.01**
Single transect - Centreline	Slope [°]	Range	9.54	10.06	1.19	13.10	2.76	18.06
		Mean	0.51	-2.42	**-0.21**	0.91	-0.46	-0.53
		Stdev	0.57	-1.31	-0.56	1.06	-0.58	0.58
	Position [-10 - 10]	Range	6.83	5.89	8.62	**2.29**	4.53	5.58
		Mean	**-0.01**	0.81	-0.25	**-0.04**	-0.21	0.37
		Stdev	**0.02**	-0.10	0.42	**-0.01**	0.06	-0.06
	Aspect [-1 - 1]	Range	**0.00**	0.15	**0.00**	0.02	0.01	**0.00**
		Mean	-0.21	**-0.01**	-0.38	-0.19	0.12	-0.10
		Stdev	0.08	0.01	-0.05	0.08	0.09	0.02

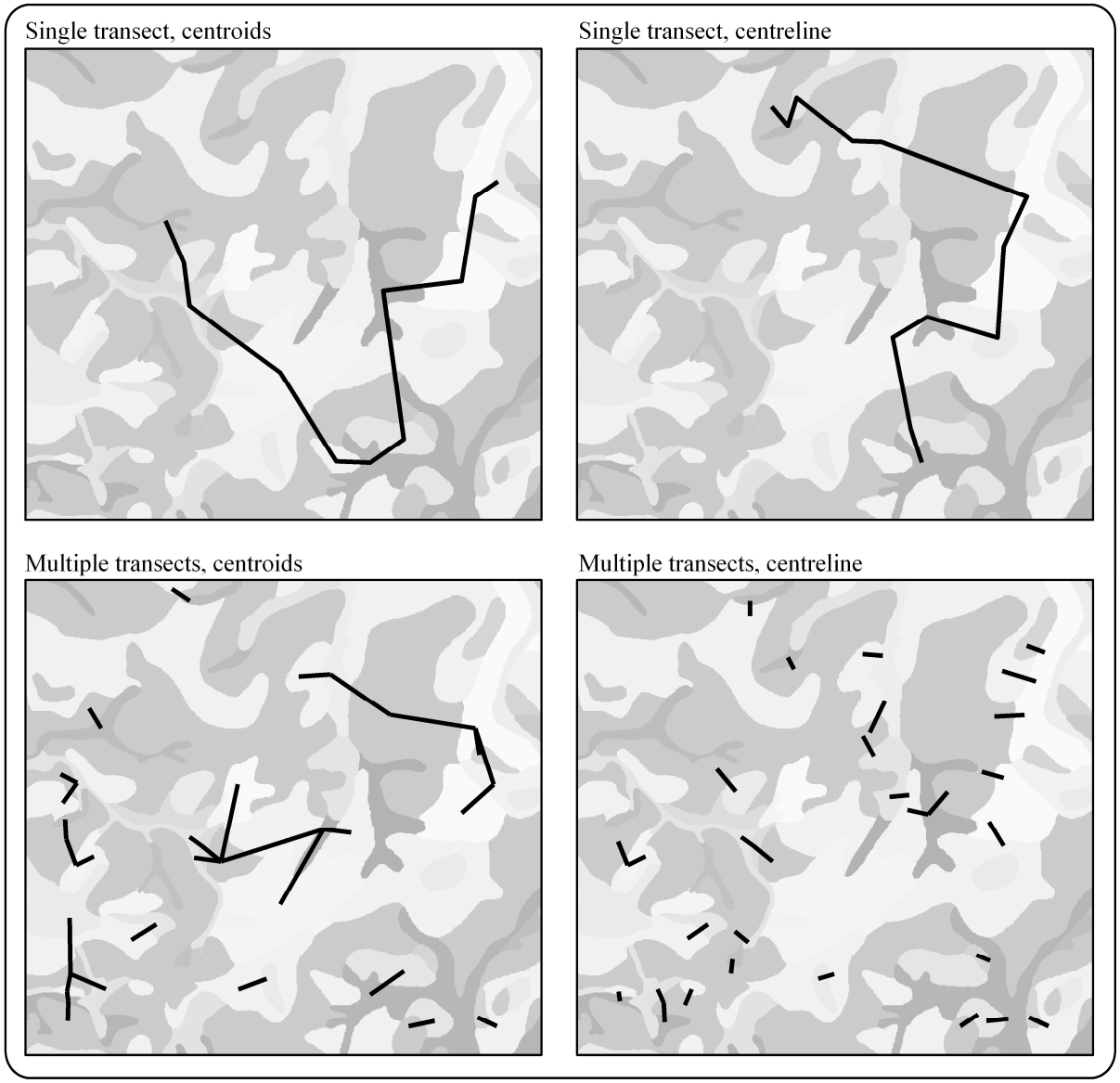


Fig. 6. Location of the transects based on the different approaches introduced for the Lower Vogelsberg (VL).

5 Conclusions

The concepts presented in this methodological study allow the generation of representative shortest transects based on class-area maps. The aim was to provide sampling schemes for linear operated proximal soil sensing techniques like electro-magnetic induction, ground penetrating radar, magnetics, near infrared spectrometry etc. in an objective and highly automated fashion.

Generally, two different approaches were introduced: a single concatenated transect and a multiple separated transects concept to cover all relevant mapping units and transition zones

respectively. Based on the aim of a soil sensing survey, the approaches can be chosen in terms of transect length which will result in a single transect, or in terms of a more complete description of the underlying environmental feature space which will be obtained by using the multiple transects approach.

Additionally, two methods were introduced to provide the nodes of a transect. For the first, one centroid was calculated per class-area. For the second, multiple points had been calculated based on the centreline of each polygon. In terms of transect length, the centreline approach must be recommended.

Concluding, we recommend to apply a combination of the multiple transects and the centreline of points approaches. In terms of a trade-off between feature space description and transect(s) length the single transect centreline approach might be chosen as it is generally shorter. However, we think that removing some transects between map units with a small adjacency of the multiple transects approach is more appropriate since the application of the multiple transects approach will and enhance informational value which should have priority. The additional travel time between the transects will partly be compensated with reduced computation time and complexity especially for large datasets. Another advantage of the multipe transects approach over the single transect approach is that is has a lower spatial autocorrelation between the residuals (Hengl et al., 2003).

The approaches introduced might help to refine the resolution for medium to large area DSM approaches based on proximal soil sensing techniques. Furthermore, the multi transects approach might help to get a better understanding of important soil-association transition zones in a landscape and can additionally be used as a tool for boundary quality estimations of existing soil maps.

Comparative high resolution digital soil mapping and sensing approaches are needed to validate the outcome in terms of prediction accuracy and resolution for large areas based on the different shortest representative transect concepts as well as other approaches like the k-th order random toposequences approach as introduced by Odgers et al. (2008).

Acknowledgements

The Collaborative Research Centre (SFB) 299 of the German Research Foundation (DFG) provided partial funding for this work. We would like to thank the Hessian Agency for Environment and Geology for providing the data.

References

Acres, B. D., Green M. A., Rackham, L. J., 1993. A method for identifying soil catenas and determining map unit composition used in a reconnaissance soil survey in Tanzania. Geoderma, 57(4), 387-404.

Behrens, T., Scholten, T., 2006. Digital soil mapping in Germany – a review. J. Plant Nutr. Soil Sci. 169, 434-443.

Behrens, T., Schneider, O., Lösel, G., Scholten, T., Hennings, V., Felix-Henningsen, P., Hartwich, R., 2008a. Analysis on pedodiversity and spatial subset representativity - the German soil map 1:1,000,000. J. Plant Nutr. Soil Sci., accepted.

Behrens, T., Schmidt, K., Zhu, A-X., Scholten, T., 2008b. Multi-scale digital terrain analysis and feature selection in digital soil mapping. submitted.

Behrens, T. Schmidt, K., and Scholten, T., 2008. An approach to removing uncertainties in nominal environmental covariates and soil class maps. In: Hartemink, A., McBratney, A., Mendoca-Santos, M.L.,: Digital Soil Mapping with Limited Data. Springer.

Burnham, K. P., Anderson, D. R., Laake, J. L., 1980. Estimation of density from line transect sampling of biological populations. Wildlife Monographs, 72, 202pp.

Bushnell, T.M., 1942. Some aspects of the soil catena concept. Soil Science Society of America Proceedings 7, 466-476.

de Zorzi, P., Barbizzi, S., Belli, M., Ciceri, G., Fajgelj, A., Moore, D., Sansone, U., van der Perk, M., 2005. Terminology in soil sampling. Pure and Applied Chemistry, 77, 827–841.

Dijkstra, E. W., 1959. A note on two problems in connexion with graphs. Numerische Mathematik. 1, 269–271.

Drummer, T. D., McDonald, L. L., 1987. Size bias in line transect sampling. Biometrics 43, 13-21.

Gerber, R., Felix-Henningsen , P., Behrens, T., Scholten, T., 2007. A new approach to detect Pleistocene Periglacial Slope Deposits via Ground Penetrating Radar as a tool for non-destructive soil depth mapping. Journal of Applied Geophysics. submitted.

Gessler, P.E., Moore, I.D., McKenzie N.J., Ryan P.J., 1995. Soil-landscape modeling and spatial prediction of soil attributes. Special issue: Integrating GIS and Environmental Modeling. International Journal of Geographical Information Systems, 9(4), 421-432.

Gobin, A., Campling, P., Deckers, J., Feyen, J., 2000. Integrated toposequence analyses to combine local and scientific knowledge systems. Geoderma 97, 103-123.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon contents on Barro Colorado Island - Digital soil mapping using Random Forests analysis. Geoderma, 146, 102-113.

Gutin, G., Punnen, A. P., 2006. The traveling salesman problem and its variations. Springer.

Hedley, S. H., Buckland, S. T., 2004. Spatial models for line transect sampling. Journal of Agricultural, Biological and Environmental Statistics, 9(2), 181-199.

Hengl, T., Rossiter, D.G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Australian Journal of Soil Research, 41, 1403-1422.

Hoosbeek, M.R. and Bryant, R.B., 1992. Towards the Quantitative Modeling of Pedogenesis - A Review. Geoderma, 55, 183-210.

Jenny, H. 1941. Factors of soil formation. McGraw-Hill, New York. 109p.

Jensen, A.J., 1996. Subsampling with line transects for estimation of animal abundance. Environmetrics, 7(3), 283-289.

Kruskal, J.B., 1956. On the shortest spanning subtree and the traveling salesman problem. In: Proceedings of the American Mathematical Society, 7, 48–50.

Lillesand, T.M.; Kiefer, R.W., 2000. Remote Sensing and Image Interpretation. John Wiley and Sons, New York.

Manly B.F.J., 2002. Estimating a resource selection function with line transect sampling. Journal of Applied Mathematics and Decision Sciences, 6(4), 213-228.

McBratney, A.B., Mendonça-Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117. 3-52.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbara, M.S., Shatara, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97(3-4), 293-327.

McKenzie, N.J., Austin. M.P., 1993. A quantitative Australian approach to medium and small scale survey based on soil stratigraphy and environmental correlation. Geoderma 57:329–355.

McKenzie, N.J., Ryan. P.J. 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94.

Milne, G., 1935. Normal erosion as a factor in soil profile development. Nature, 138: 548-549.

Odgers, N., McBratney, A., Minasny, B., 2008. Generation of k-th order random toposequences. Computers and Geoscience. In Press.

Schmidt, K., Behrens, T., Friedrich, K., Scholten, T., 2008. An approach to construct soilscapes by segmenting soil maps for digital soil sensing and mapping in homogeneous feature spaces. submitted.

Sommer M., Schlichting E., 1997. Archetypes of catenas in respect to matter - A concept for structuring and grouping catenas. Geoderma 76, 1-33.

Stoyan, D., 1982. A Remark on the Line Transect Method. Biometrical Journal, 24(2), 191-195.

Viscarra-Rossel, R., Taylor, J.A., McBratney, A.B., 2007. Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. European Journal of Soil Science, 58(1), 343-353.

Webster, R., Oliver, M.A., 1990. Statistical Methods in Soil and Land Resource Survey. Oxford University Press, Oxford.

Manuscript 3

Instance selection and classification tree analysis for large spatial datasets in digital soil mapping

Geoderma (2008), 146: 138-146

Karsten Schmidt, Thorsten Behrens, and Thomas Scholten

Institute of Geography, Chair of Physical Geography, Eberhard Karls University

Tübingen, Rümelinstraße 19-23, D-72074, Tübingen, Germany

Abstract

Digital soil mapping is currently experiencing a tremendous increase in available environmental covariates and resolution for spatial soil predictions, resulting in computational problems in terms of limited data handling capabilities of machine learning approaches. This is of particular importance when gridded spatial soil class maps are used as a basis for predictions containing large amounts of redundant instances and noisy information.

In this study we systematically analyze the effect of instance selection, which aims at reducing sample size, while preserving or even increasing prediction accuracy. On a soil class dataset with 95.000 instances we tested two sampling approaches in relation to parameter settings of decision tree based learning: proportional and disproportional stratified random sampling. An automated grid search approach was used to find the best performing parameter settings of the decision tree.

The results show that an appropriate sampling method in combination with a grid search method returns better results than those obtained when grid learning is applied without instance selection. Instance selection increases prediction accuracy especially if the frequency distribution of the soil classes is low compared to the surrounding area. However, instance selection does not help in pedological interpretation. Nevertheless, it is a valuable pre-processing method to handle large spatial high resolution datasets in digital soil class prediction in terms of accuracy and computational costs.

As suggested on the basis of the results of this study, spatially constrained instance selection as well as boundary based digital soil mapping in terms of soil taxonomic contrast should be investigated in future pedometric research.

1 Introduction

Within the last decade an increase in the availability of explanatory variables to solve the function of soil forming factors (Jenny, 1941), reformulated for quantitative empirical modelling by McBratney et al. (2003), can be observed and is expressed in a dramatic increase of research publications in this field of soil science (McBratney et al., 2000; 2003). This is promising, as the global lack in the availability of soil data can now be counteracted using digital soil mapping approaches. However, data handling becomes more complex due to the high amount of available potential predictors as well as the large amount of pixels when existing rasterized soil maps are used as a basis for extrapolation. Thus, the lack of predictors is now replaced by the limitation of the data handling capabilities of the algorithms used for prediction, which limits the analysis of soil formation and distribution.

One of the most widely used and best performing inductive learning algorithms in terms of generating interpretable rules as well as prediction accuracy are classification tree algorithms (Breiman et al., 1984; Loh and Vanichsetakul, 1988; Mitchie et al. 1994; Behrens and Scholten, 2006b). Classification trees are non-parametric (Friedman, 1991; Mitchie et al. 1994), non-sensitive to the presence of missing data and to the inclusion of a large number of irrelevant features (Schafer, 1997; Hastie et al., 2001), and are described as a robust prediction technique (Loh and Vanichsetakul, 1988; Lagacherie et al., 2001; Scull et al., 2005). Applications in environmental sciences can thus be found in various disciplines like ecology (e.g. Geng et al., 2004; Munoz and Felicisimo, 2004), remote sensing (e.g. Hansen et al., 1996; Friedl and Brodley, 1997; Debeir et al., 2001; Gómez-Chova et al., 2003; He et al., 2003) and soil science (e.g. Lagacherie and Holmes, 1997; Zhang et al., 1999; Bui et al., 1999; Giasson et al., 2000; Moran and Bui, 2002; Bui and Moran, 2003; Zhou et al., 2004; Behrens and Scholten, 2006a,b; Geissen et al., 2007).

Yet, handling large datasets using decision trees can be inefficient in terms of learning time and prediction accuracy due to redundant and noisy information and often results in complex models (Liu and Motoda, 1998; Brighton and Mellish, 2002). This is particularly true when using gridded spatial datasets as a basis for predictions (Munoz and Felicisimo, 2004). In contrast to datasets originating from point sampling schemes, mostly counting only up to a view hundred samples, digital soil class prediction based on existing soil class maps can easily comprise several thousands to millions of training “samples” if the prediction is based on gridded spatial datasets where every pixel serves as a separate sample (Bui et al., 1999; Shrestha et al., 2004; Behrens et al., 2005). This leads to large amounts of redundant information, causing negative effects not only with regard to computation time but also to prediction accuracy (Qi, 2004). Lagacherie and Holmes (1997) showed that inconsistencies within the training dataset, such as noise, can greatly influence predictive accuracy.

To handle datasets containing redundant and/or noisy instances (samples) as well as multi-collinearity two main branches within statistical learning research do exist: instance selection (Liu and Motoda, 2001) and feature selection (John et al., 1994).

Feature selection aims at reducing the feature space to the driving factors and at reducing multi-collinearity. In contrast, instance selection is applied to reduce the dataset by fitting out the relevant samples. The application of feature and/or instance selection depends on the structure of the dataset and the learning approach used. As the dataset used in this study contains large amounts of samples, and as decision trees are robust to correlated features, we focus on instance selection.

In pedometric research the discussion on instance selection has been limited. Even though some methods of classical spatial soil sampling designs (Cochran, 1977; Brus, 1993; Brus and DeGruijter, 1997; Domburg et al., 1997) and instance selection (Gu et al., 2001) are based on the same theoretical concepts, the aim of instance selection is contrary to the aim of soil sampling. In soil sampling it is most important to optimize sampling schemes to derive a sample set that is as sparse as possible to save labor and lab costs. In instance selection, more than enough samples are available. The challenge here is to extract a representative subset that is still large enough that no relevant information gets lost but is small enough that it can be handled easily by learning algorithms.

Moran and Bui (2002) were the first to examine instance selection in a digital soil mapping approach by comparing two random sampling methods applied over all soil classes of the entire training dataset. In this study we go a step further in systematically analyzing instance selection on the basis of single soil classes, which is important to compare the outcome of different spatial sample distributions and relations to soil forming factors. Another approach reported by Qi (2004) addresses the task of noise reduction by selecting samples based on fitted histograms of the predictors. However, this approach is not applicable for large feature spaces (many predictors) as each feature has to be analyzed separately.

In this study we combine instance selection with an investigation of reasonable parameter settings for classification trees. The analysis of interactions between sample sizes, sample method, and tree parameters aims at stable digital soil mapping models with reduced complexity, faster computation times, as well as easier interpretability.

2 Rationale

In this study we systematically compare different random sampling designs and sample sizes to analyze the impact of instance selection on a digital soil class mapping task. As different model parameters of decision tree algorithms are directly related to sample size, i.e. the minimum

samples in an end node as well as the cross validation pruning settings (Breiman et al., 1984; Mitchell, 1997), it is important to analyze these parameter settings in relation to instance selection approaches. In order to automatically test different parameter settings to find optimized model parameters, a so called grid search (grid- or hyper-learning) approach (cf., Gourieroux and Monfort, 1995; Gilardi and Bengio, 2001; Jin, 2006) is applied. In grid search procedures different parameter settings are systematically analyzed in p-dimensional parameter spaces (cf. Chapter 3.3).

The approach presented in this paper is illustrated in Fig. 1 and organized as follows:

- First, biggest trees (no cross validation pruning) are calculated as references to analyze the effect of overfitting (Behrens and Scholten, 2006b) and for the subsequent analysis on parameter and instance selection settings.
- Second, different cross validation pruning settings are tested to analyze different parameter settings for the entire dataset to understand the behavior of the tree models in terms of overfitting and to obtain an additional reference to interpret the effect of instance selection.
- Third, different instance selection approaches and sample sizes are analyzed on the basis of the same grid search approach as applied on the entire dataset to examine the impact on prediction accuracy.

All approaches are computed for binary (2-class) predictions, i.e. each soil class is tested individually to gain a deeper insight about the possible variations of results and as a basis for pedological interpretations. Furthermore, tree complexity in terms of the number of end nodes is discussed.

Fig. 1. Overview of methods.

3 Materials and methods

3.1 Study area and datasets

The investigation area, a low mountain range within the south-west German-Lorraine Triassic escarpment (Rhineland-Palatinate; Germany), comprises 350 km^2 (Fig. 2).

The prediction approach is based on 40 terrain attributes derived from a digital elevation model with a resolution of 20 m (Behrens et al., 2005; Behrens and Scholten, 2006b). The local soils have formed from substrates of Upper Red Bed Sandstone, Bunter and Lower Limestones. The training as well as the validation area, each comprising 40 km^2, contain the same 6 soil classes (SC1 - SC6) mapped at a scale of 1: 50.000 (Table 1). According to the DEM the soil map was rasterized to a resolution of 20 m. To compare the training with the validation area Fig. 3 shows feature spaces for elevation, slope, aspect, and curvature. As differences in feature spaces are marginal this setting serves as an optimal test bed for induction based learning (Behrens and Scholten, 2006b).

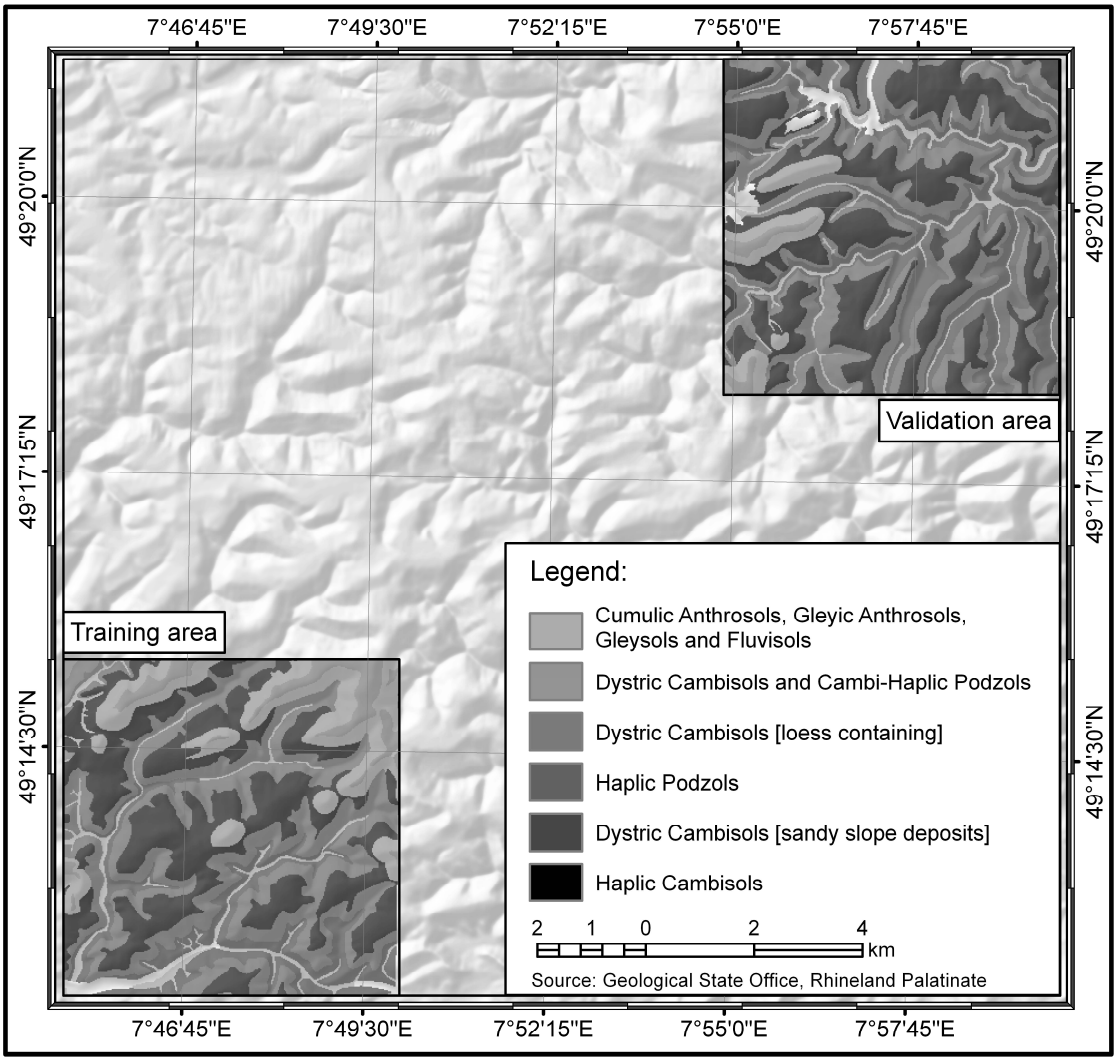


Fig. 2. Location of the investigation area in the Palatinate Forest, Germany.

As data mining based predictions strongly rely on predictors and their feature spaces, predictions can only reasonably be extrapolated to areas where the same features with a similar feature space are available, i.e similar soilscapes (Hole, 1978; Lagacherie et al., 2001).

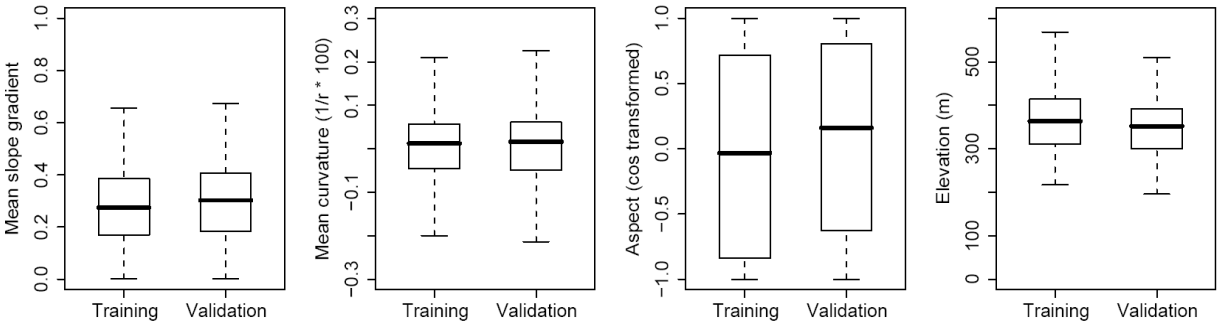


Fig. 3. Feature spaces of slope, curvature, aspect, and elevation of training and validation area.

As the study area is located in a low mountain range, relief plays an important role in pedogenesis in terms of erosion and accumulation. However, due to differences in parent material as well as a stratification with almost no dip, geology has a strong impact on soil formation in this region. Thus, including information on parent material in the prediction approach will lead to very high accuracies that will make the differences in prediction and instance selection approaches hard to interpret. Additionally, in terms of a "real world" example using digital elevation models only is more realistic as digital elevation models are widely available whereas large to medium scale geological maps are not (Behrens et al., 2008). Hence, geological information was not included in this study as well as other spatial predictors. Other ancillary variables like remote sensing or climate data were not available within this study.

Table 1

Description and coverage of the 6 soil types according to World Reference Base for Soil Resources (WRB) (IUSS Working Group, 2006) and the corresponding reference ID.

Soil classes	ID	Coverage [%]
Haplic Cambisols	SC1	33
Dystric Cambisols and Cambi-Hablic Podzols	SC2	13
Dystric Cambisols [loess containing]	SC3	5
Haplic Podzols	SC4	33
Dystric Cambisols [sandy slope deposits]	SC5	12
Cumulic Anthrosols, Gleyic Anthrosols, Gleysols, and Fluvisols	SC6	4

3.2 Instance selection

Instance selection is a technique to reduce the size of a dataset by extracting relevant information, where the new subset resembles the original dataset in further analysis (Pal and Jain,

2005). Instance selection is generally applied for three technical reasons: *enabling*, *focusing*, and *cleaning* (Liu and Motoda, 2002) described below:

- *Enabling*: As the capability to handle large datasets is limited for every data mining algorithm, instance selection enables these algorithms to function and work effectively (Moran and Bui, 2002; Grinand et al., 2007; Behrens et al., 2008).
- *Focusing*: Soil maps are generally not intended for the purpose of applying data mining techniques (Qi, 2004). For example, traditional soil mapping is based on a mental model by a soil surveyor (Bui and Moran, 2003; Bui, 2004) focused on a detailed soil profile description to subsequently construct a map of soil classes at a specific scale. These soil classes generally comprise larger regions with some hundreds to thousands of pixels of an underlying DEM resulting in highly redundant information. Thus, focusing aims at selecting the relevant information.
- *Cleaning*: No soil map is perfect in terms of decision boundaries as well as spatial delineation (Burrough and McDonnell, 1998; McKenzie and Ryan, 1999; Zhou et al., 2004; Behrens et al., 2008). A saying in computer science often used in the context of noise in datasets is "garbage-in-garbage-out". In this respect, cleaning a dataset helps removing, or at least reducing noisy data (Qi, 2004; Behrens et al., 2008). In contrast to that, high quality soil data will lead to high quality results and reduced computational and labor costs.

Summarizing, the main aim of instance selection is to reasonably reduce large datasets for faster predictions while preserving accuracy (Liu and Motoda, 1998; Bui et al., 1999). Hence, the ideal outcome of instance selection is model independent and can be described as follows:

$$P(MeD) = P(MsD), \quad (1)$$

where P(MeD) is the predictive power of a model based on the entire dataset and P(MsD) is the predictive power of a subset (Liu and Motoda, 2001). On the one hand, in an optimistic case, "less is more" (Liu and Motoda, 1998), so that the resulting prediction accuracy based on the subset is higher than the one for the entire dataset. On the other hand, instance selection can lead to a trade-off between sample size and mining quality (Brighton and Mellish, 2002).

In this study on soil class prediction we apply proportional stratified random sampling (PS) and disproportional stratified random sampling schemes (DS) as instance selection approaches (Gu et al., 2001). The soil-sampling counterparts of these two approaches are well-known in spatial soil sampling design aiming to improve estimates (Brus, 1993, Domburg et al., 1997).

Proportional stratified random sampling accounts for the frequency distribution of each soil class in the entire dataset. In the disproportional approach the same amount of instances is

selected for all classes. This approach is recommended by Kohonen et al. (1995) for supervised classification applications, even when the a-priori probabilities are skewed.

We compare the original dataset comprising 95.000 pixels with 6 different subsets sizes of 500, 1000, 2500, 5000, 7500 and 10.000 samples for both sampling approaches. For each sample size three independent iterations were carried out to provide information on sampling variance. The results are discussed on the average of these three predictions.

3.3 Prediction and validation

3.3.1 Classification tree

Classification tree analysis is a supervised non-parametric statistical classification approach based on binary recursive partitioning techniques (Breiman et al., 1984; Friedman, 1991; Loh and Shih, 1997). The step-wise constant partitioning scheme (Friedman, 1991) provides increasingly homogeneous subsets in terms of the dependent variable – in our case the soil classes - based on existing observations. It is used to identify rules which can subsequently be applied for extrapolation. Partitioning is stopped if a minimum tolerated amount of samples (pixels) in a node of the tree is reached. This threshold influences the size of the tree in terms of end nodes and is thus strongly related to overfitting and generalization (Mitchie et al., 1994). For each terminal node the majority class label is assigned for the final classification results.

To construct a classification tree we apply the CRUISE algorithm as introduced by Loh and Shih (1997). Compared to the greedy search approach of CART (Breiman et al., 1984), CRUISE is based on an analysis of variance which is reported to be stable and unbiased in terms of variable and split point selection (Kim and Loh, 2001).

The 1D CRUISE - algorithm used in this study separates the tree growing process into two steps: first selecting the split variable and second selecting the split point. To select the most important variable for each split an ANOVA F-Test is computed. The variable with the smallest p-value, which determines the significance of the features, is chosen to separate between two classes. Once the variable with the smallest p-value is selected, a threshold that defines the split point is computed applying a linear discriminant analysis (LDA) (Kim and Loh, 2001). As LDA is most effective when the data are normally distributed with the same covariance matrix, a Box-Cox transformation is calculated to adjust the response variable (cf. Qu and Loh, 1992).

3.3.2 Validation

Predicted categorical spatial data are validated using a confusion matrix (van Rijsbergen, 1979; Ishioka, 2003) by deriving measures of effectiveness such as recall, precision and the F-measure

which are important measures for information retrieval (van Rijsbergen, 1979; Raghavan et al., 1989; Manning and Schütze, 1999; Giasson et al., 2000; Zhu, 2000; Behrens et al., 2005). Table 2 shows a confusion matrix.

Table 2

Confusion matrix

		Original soil unit	
		True	False
Predicted soil unit	True	tt	tf
	False	ft	ff

Recall (rc) describes the relation between positive (tt) and negative predicted pixels (ft) and thus the probability that a mapped soil class sample (pixel) is actually predicted, i.e. it represents underestimation:

$$rc = \frac{tt}{tt + ft}, \tag{2}$$

whereas precision (pc) describes the probability that a predicted soil class pixel is actually mapped, i.e. it represents overestimation:

$$pc = \frac{tt}{tt + tf}. \tag{3}$$

To quantify the overall model quality in a composite measure the F_1-measure (van Rijsbergen, 1979), representing the harmonic mean of under- and overestimation, is frequently used. It is calculated as follows:

$$F_1 = \frac{2 * rc * pc}{rc + pc}. \tag{4}$$

3.4 Grid-search

Grid-search or hyper-learning (Ensor and Glynn, 1997; Bergez et al., 2004) is a simple, yet time consuming procedure to find the best parameter settings or combinations for fitting a model in a *p*-dimensional parameter space systematically. Wrapped around a decision tree algorithm it leads to *p* trees based on *p* different parameter settings. The most important parameters controlling the tree size as well as overfitting, which consequently influence validation accuracy,

are the minimum data in a terminal node (mindat) and the standard error (SE) of a 10-fold cross validation pruning (Breiman et al., 1984). Thus, we apply a 2D grid learning scheme. Based on the size of the soil dataset, the F_1-measure was calculated for each combination of a mindat of m = 5, 10, 50, 100, 500, 1000, 5000, 10.000 and the entire training set size as well as SE-values of SE = 0, 1, …, 10. Additionally, un-pruned (biggest) trees based on a mindat of 5 were calculated as reference models for evaluating each combination of mindat and SE.

3.5 Model complexity

As an additional basis to interpret the model performance, the results of the instance selection approaches as well as the impact of noise and redundancy on the prediction algorithm we investigate model complexity. With regard to tree based approaches the number of end nodes can be used as a simple measure which is mainly determined by three parameters: sample size, mindat, and the cross validation pruning settings. The more samples and the smaller mindat the larger a tree can grow. The more it overfits this way (i.e. learns noise instead of relevant information) the more it will be reduced by the subsequent pruning.

The more end nodes in a tree the more complex the model and the more complex the interpretation of the results (Munoz and Felicisimo, 2004).

We compare the model complexities of all approaches: biggest trees, grid search over the entire dataset, as well as the instance selection approaches with their three underlying random sampling iterations.

4 Results and discussion

4.1 Classification tree analysis

4.1.1 Biggest tree reference predictions

The analysis of the biggest tree models for each soil class shows varying results in terms of predictability (Table 3). Soil classes SC2, SC3 and SC5 show prediction accuracies in the validation area below 0.5 which is due to two major reasons: First, they show a weak relation to relief, and second they contain noisy information in terms of imprecise delineations, resulting from a high degree of taxonomic similarity of soils mainly differentiated by altering Loess contents.

The comparatively high prediction accuracy for these soil classes in the training area serves as a first indicator for overfitting, due to too complex tree models based on noisy data.

Table 3 Prediction accuracy (F_1) and number of terminal nodes (nodes) for the biggest tree settings (BT, no cross validation pruning, minimum instances in the end node - mindat = 5), for the optimized tree model parameters SE (Standard Error) and mindat based on a grid search method (GS), for the optimized settings for the tree model parameters (SE and mindat) and sample sizes (S) for proportional stratified random sampling (PS) and disproportional stratified random sampling (DS) averaged over three independent runs.

		SC1	SC2	SC3	SC4	SC5	SC6
BT	F_1T	0,81	0,80	0,77	0,73	0,66	0,69
	F_1V	0,72	0,41	0,25	0,52	0,43	0,55
	Nodes	622	358	265	625	477	344
GS	SE	1	0	9	4	2	0
	Mindat	100	5	100	10	1000	5000

	F_1T	0,78	0,81	0,65	0,67	0,50	0,52
	F_1V	0,72	0,42	0,28	0,52	0,48	0,57
	Nodes	62	23	14	45	28	5
DS	SE	4	0	0	1	1	2
	Mindat	5	5	5	10	5	5
	S	10.000	7500	10.000	5000	7500	7500
	F_1T	0,77	0,73	0,65	0,70	0,58	0,50
	F_1V	0,78	0,36	0,19	0,62	0,61	0,54
	SD_T	0,004	0,017	0,012	0,003	0,004	0,008
	SD_V	0,019	0,010	0,005	0,008	0,005	0,003
	Nodes	30	126	145	47	58	21
PS	SE	0	0	1	0	2	0
	Mindat	1000	5	5	10	50	500
	S	10.000	10.000	10.000	10.000	10.000	7500
	F_1T	0,73	0,74	0,68	0,70	0,54	0,52
	F_1V	0,75	0,40	0,31	0,54	0,47	0,58
	SD_T	0,007	0,029	0,029	0,009	0,004	0,0001
	SD_V	0,005	0,026	0,018	0,008	0,023	0,001
	Nodes	14	48	24	60	22	2
	ΔF_1V [%]	5,49	-2,45	1,24	9,86	12,90	0,53

Based on the optimized settings in sample size, mindat, and SE the variation of the reiterations is expressed through the standard deviation (SD) for both training (T) and validation (V) area.
The information gain for the optimal instance selection method compared to parameter optimization (GS) was finally calculated by the difference (ΔF_1V [%]) between the single validation results, thus positive values indicate better results for instance selection, negative values indicate better results for parameter optimization.

The biggest tree results serve as primary reference measures to evaluate the validation accuracy obtained by parameter fitting and/or instance selection.

4.1.2 Grid search

The prediction results over the different settings for mindat and SE as analyzed by the grid search approach are shown in Fig. 4, revealing varying impacts to the different soil classes.

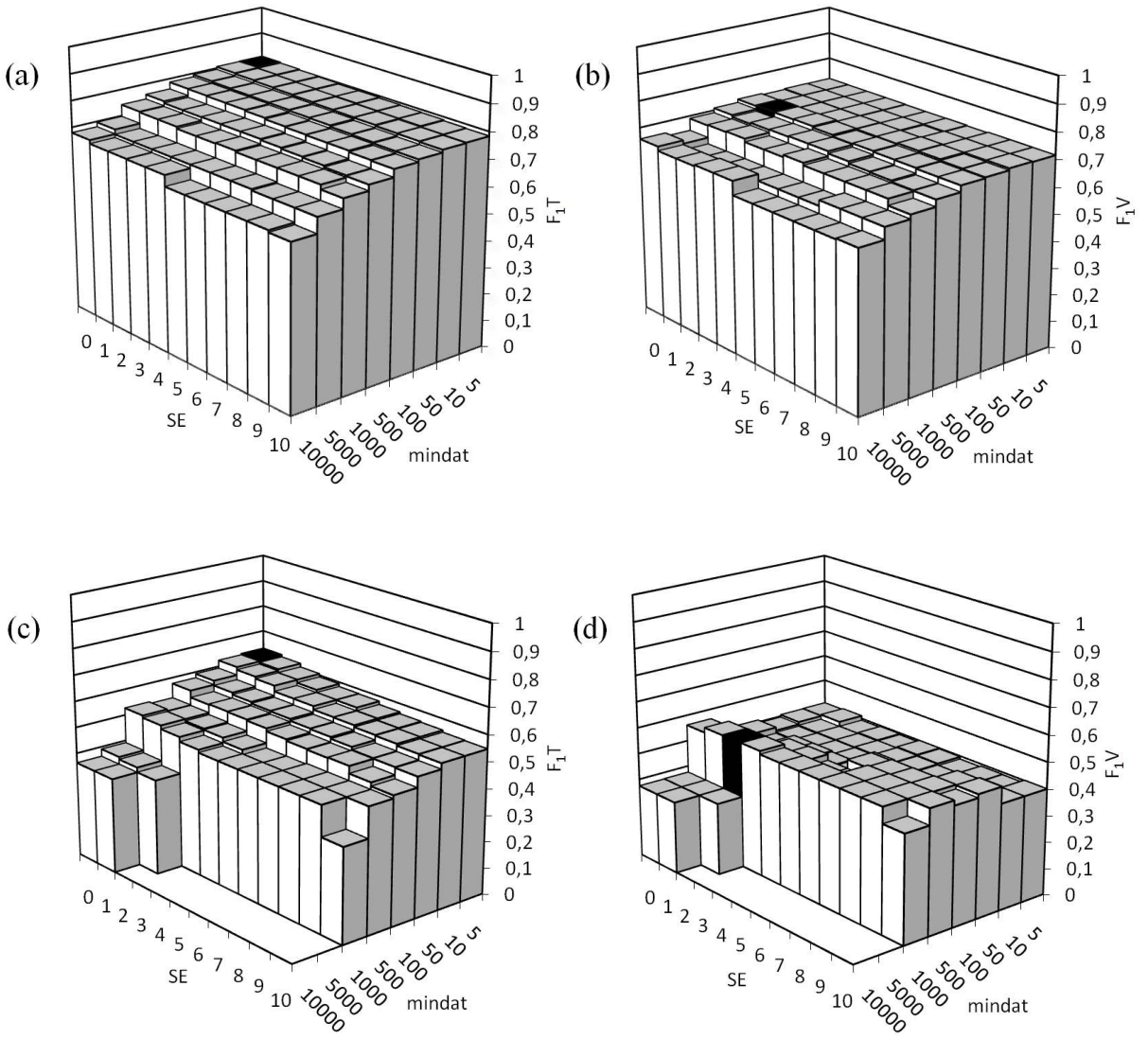


Fig. 4. Prediction accuracy (F_1) for the training (T) and validation (V) area of different model settings for SC1 (a, b) and SC5 (c, d). The best results are marked in black.

It can be seen that SC 1 produces relatively stable results across the parameter space of the grid-learning scheme, whereas SC 5 shows unstable results. As SC1 can be extrapolated accurately into the validation area (Fig. 4b) it seems to be highly correlated to relief and accurately mapped, whereas SC5 shows a decrease in prediction accuracy of 11 % between training and validation area. In this case, less complex models in terms of higher mindat return better results in the validation area.

The results shown in Table 3 reveal a high variability of the optimal model settings over the soil classes. The differences between the optimized settings are too high to derive optimal global

settings, which might be applied for all soil classes as almost the entire range of mindat and SE occurs. Generally, the differences in prediction accuracy between the soil classes can be related to missing predictors (e.g. geology) or noise in terms of mapping precision (Fig. 5).

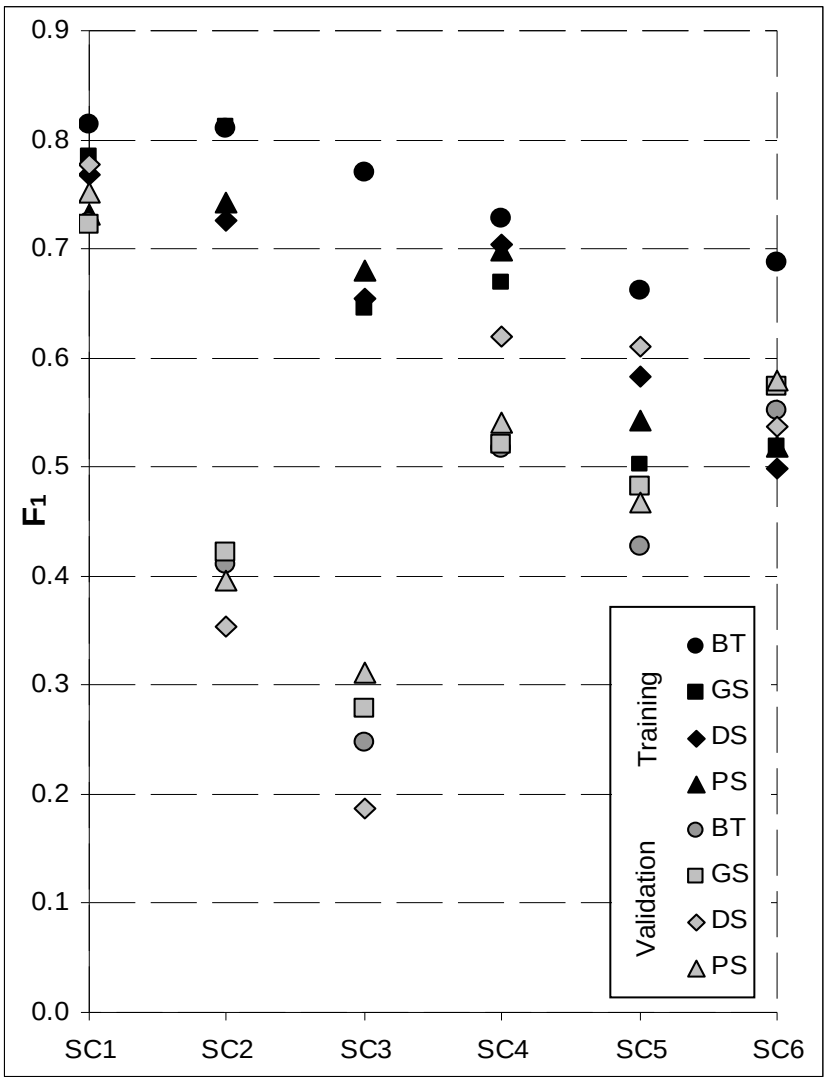


Fig. 5. Optimized prediction results for disproportional stratified random sampling (DS), proportional stratified random sampling (PS) and grid search based optimized parameter settings (GS) for each soil class compared to the biggest tree results (BT) for both training and validation area.

Grid search offers the possibility to analyze the effect of overfitting. Generally, overfitting is indicated by large differences between the prediction accuracy in the training and validation areas. Table 3 shows that for all soil classes except SC2 the accuracy in the training area decreases due to the grid search approach. In no case the validation accuracy decreases. Thus, the generalization rate is higher for all soil classes except for SC2. Concerning SC2 it is remarkable that the training accuracy increases even though the computed model is more than 15 times less complex in terms of terminal nodes.

Even if the tree size is generally expected to correlate with the amount of samples provided (Munoz and Felicisimo, 2004), the values for mindat und SE cannot be related to sample size and also for larger soil classes higher values for mindat cannot be found (Table 1 and Table 3).

4.2 Instance selection

The optimized sampling sizes for the instance selection approaches range from 5000 up to 10.000 pixels (Table 3). Values below 5000 return lower validation accuracies for all soil classes, indicating that relevant information gets lost. Most remarkably - except for the F_1-measure in the validation area of SC2 - at least one of the two sampling approaches returns better results as obtained by the grid learning approach for the entire dataset. For some soil classes (SC4, SC5) the effect of instance selection boosts the validation accuracy up to 13%.

The general relation between sample size and model parameters can be seen by analyzing the optimized values for mindat, which are lower for the reduced sample sets, based on the instance selection approaches than for the grid search approach based on the entire dataset (Table 3). The only exception is SC1-based proportional stratified random sampling, resulting from different settings of the standard error, which are typically higher for the grid search approach based on the entire dataset.

The differences between the instance selection methods PS and DS are partly related to the frequency of the soil classes, such that smaller soil classes return better validation results when proportional stratified random sampling is applied. This seems contrary to what is expected as less pixels of a small soil class are sampled compared to its surrounding area (Chen et al., 2004).

The high positive impact of instance selection in digital soil class mapping with an average increase of 7% for the F_1-measure in the validation area is shown in Table 4.

This positive effect holds true for all three independent random selections carried out for every prediction in this study, which is shown by low standard deviations in Table 3.

The average variation in prediction accuracy of the three random sampling iterations for the optimized DS settings is 1.6 % in the training area and 1.8 % in the validation area in terms of the F-measure. For PS the corresponding values are 4 % and 3 %. Thus, variation does not increase in the validation area.

Table 4

Average prediction accuracy (F_1) of the approaches tested (T = training area, V = validation area, BT - biggest tree, GS - optimized parameter settings based on a grid search method, S - optimized sampling schemes)

Method	F_1T	F_1V

BT	0,75	0,48
GS	0,66	0,50
S	0,67	0,55

As the variation of the three independent PS samplings is within the range of the average improvement in accuracy achieved for SC2, SC3, and SC6, it can be stated that PS is not very useful in terms of increasing accuracy compared to grid learning solely (Table 3). Nevertheless, it speeds up computation.

Concerning DS, which is superior to PS for SC1, SC4, and SC5, variation is smaller than for PS and - more important - is below the rate of increase in accuracy (Table 3). Thus, for these soil classes DS is recommend both for faster computation and for increasing prediction accuracy. This confirms the finding of Kohonen et al. (1995) who are recommending DS.

Yet, for the some units with the lowest proportion (SC3, SC6) PS provides up to 12 % better results than DS. Thus, the recommendation of Kohonen et al. (1995) is not true for extremely skewed a priori proportions or imbalanced data. This is difficult to explain, as boosting soil classes with small proportions via sampling should produce better results at first sight (Chen et al., 2004). An explanation might be that important information of spatially neighbouring regions gets lost (Chen et al., 2004). This is especially the case if soil classes are mapped imprecisely and more information about spatially neighbouring regions is needed to average the noise between the soil class and its surrounding and thus to increase the ability of the model to generalize. This is emphasized by the high mindat of 500 for the optimized PS settings of SC6 compared to a mindat of 5 for DS, indicating overfitting by a too complex model. Thus the problem seems to be correlated to noise rather than spatial coverage. This can be explained soil scientifically for the valley bottom soils (SC6). They are mapped according to their location in the geological map which is problematic because of three reasons: first, even if there are overlaps between geological maps and soil maps, the main purpose of the geological map is not to indicate soil distribution. Second, the valley bottoms in geological maps are mostly generalized and thus do not exactly fit to the location as shown in current DEMs. Third, geological maps for this region are drawn on the basis of older topographical maps resulting again in slightly different spatial delineations. Thus, predictions based on terrain attributes only return weaker results than soil classes mapped on the basis of current topographical maps.

The prediction results might also be affected by missing features. As the formation of SC2 and SC3 is strongly related to geology (Behrens and Scholten, 2006b) information on relief is not sufficient for good predictions. For these classes which produce the worst prediction results PS is again better. The explanation is the same: more information is needed in the direct neighborhood to enable the model to generalize and not to overestimate, as reported by Behrens

and Scholten (2006b).

Concluding, if the prediction works quite well and PS produces better results for skewed distributions, the soil class is mapped imprecisely and/or important predictors are missing. Thus, Kohonen et al. (1995) are right for datasets which are not skewed, do not contain noise, but all relevant features. The possibility to overcome this problem and to achieve better generalization results for DS on skewed distributions and noisy datasets might be to use a spatially constrained sampling approach, where more samples are taken near the boundary. This should be based on spatial distance density function schemes to average out the delineation noise in the soil data and therefore to increase validation accuracy. This might lead to a new paradigm in digital soil class mapping focusing on the boundaries instead of concentrating on the more homogeneous cores of the class areas. Following the concept described by Hole (1978), different types of boundaries can be characterized by the taxonomical contrast of the adjacent soil classes. In this concept of nine orders, order 9 represents a boundary separating orders of soil and order 1 represents a boundary separating soil phases. Thus, theoretically, higher order boundaries should be easier to predict as low order boundaries, due to a higher contrast which should be expressed by a higher contrast in state factors. Further studies have to prove this concept.

4.3 Model complexity

Model complexity analysis in terms of the number of end nodes offers the possibility to explain why instance selection in combination with grid search approaches produces better results than grid search over the entire dataset only.

Table 5.3 shows that in average the tree complexities obtained from the grid search approach over the entire dataset are comparable to the instance selection approaches. This is contrary to the general expectation that smaller datasets return less complex models. It can be explained with the different mindat and pruning settings and shows the high dependency between sample size and parameter settings. As random sampling is not a procedure designed to remove noise from a dataset as for example special approaches like Wilson editing (Behrens et al., 2008) and as the different randomized sampling tests returned very similar prediction results (cf. Chapter 4.2) the surprisingly good results of the instance selection approaches are not based on dataset *cleaning* as no noise is removed. As a consequence, this phenomenon must be an effect of *enabling* (cf. Chapter 3.2) in terms of removing redundant information. In this case, CRUISE must be regarded unstable in terms of redundancy. More detailed insight can be gained from the three independent random sampling tests. Here we see that the average node size for DS and PS differs. The node size for the approaches where DS returned the best results is similar for the

entire dataset, whereas the node size for the approaches where PS returned better results is larger. As PS returns better results for skewed distributions (i.e., the soil class comprises a relatively small area in the training area) it can be stated that the prediction results obtained with CRUISE are only affected by redundancy if the distributions are much skewed. Thus, it is an effect of *focusing*.

In general, even though the tree complexity compared to the biggest tree models is reduced by parameter fitting for about 94% for the entire dataset, as well as 93 % and 92 % for all DS and PS approaches, the resulting models must be regarded too complex for interpretation in most of the cases. Thus, in this study instance selection does not help easing pedological interpretation. Additional approaches to analyze feature importance might be more efficient for this purpose (John et al., 1994, Liu and Motoda, 1998).

4.4 Summarizing conclusions

This study systematically analyzes the influence of instance selection schemes and grid learning in data mining based digital soil mapping approaches. Based on 3702 tree inductions for 6 soil classes, the results can be summarized and discussed as follows:

- Biggest trees generally lead to higher prediction accuracies in the training area and lower generalization rates compared to grid learning and instance selection due to overfitting and effects of enabling and focusing.
- Optimizing sensitive parameters increases the generalization rate due to a reduced effect of overfitting which is trivial and as expected (e.g. Breiman et al, 1984).
- An appropriate sampling method in combination with grid search returns better results than obtained when grid learning is applied on the entire dataset. This enabling effect is remarkable as the results are better than the ideal outcome of instance selection where the predictive power of a model based on the entire dataset equals the predictive power of a subset (Liu and Motoda, 2001).
- Instance selection increases prediction accuracy especially if the frequency distribution of the soil classes is low compared to the surrounding area which must be regarded as an effect of *focusing* (Liu and Motoda, 2002).
- For small soil classes (skewed frequencies) proportional sampling returns better results. This effect shows that adaptive sampling in terms of handling characteristic soil distributions and frequencies differently will lead to higher prediction accuracies compared to global approaches (Moran and Bui, 2002).

- Based on the analysis of model complexity instance selection does not necessarily boost pedological interpretations.

Instance selection and grid learning are important tools in digital soil class mapping as they both help speeding up computation and increasing prediction accuracy. Yet, it is not easy to recommend a sampling scheme a priori. To handle skewed, noisy, and redundant data in digital soil class mapping based on large datasets additional approaches like Wilson editing (Wilson, 1972; Behrens et. al, 2008), latin-hypercube sampling (Carré et al., 2007), and prototype generation (Wai et al., 2001) should be tested in further studies. As suggested on the basis of the results of this study, spatially constrained instance selection as well as boundary mapping in terms of soil taxonomic contrast (Hole, 1978) should be investigated in future pedometric research.

References

Behrens, T., Förster, H. Scholten, T., Steinrücken, U., Spies, E.-D., and Goldschmitt, M., 2005: Digital soil mapping using artificial neural networks. J. Plant Nutr. and Soil Sci. 168, pp. 21-33.

Behrens, T. and Scholten, T., 2006a: Digital Soil Mapping in Germany – a review. J. Plant Nutr. and Soil Sci. 169, pp. 434 - 443.

Behrens, T. and Scholten, T., 2006b: A comparison of data-mining techniques in predictive soil mapping. In Lagacherie, P. McBratney, A.B., Voltz, M. (Eds): Digital Soil mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam. pp. 353-364.

Behrens, T., Schmidt, K. and Scholten, T. (2008): An approach to removing uncertainities in nominal environmental covariates and soil class maps. In Hartemink., A.E., McBratney, A.B. and Mendonça-Santos, M.L., (eds) 2008. Digital soil mapping with limited data. Developments in Soil Science series. (in press).

Bergez, J.E., Garcia F. and Lapasse L., 2004: A hierarchical partitioning method for optimizing irrigation strategies. Agriculture Systems, 80: 235-253.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., 1984: Classification and Regression trees. Wadsworth.

Brighton, H. and Mellish, C., 2002: Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery 6:153-172.

Brus, D.J., 1993: Incorporating models of spatial variation in sampling strategies for soil. Doctoral thesis. Wageningen. Netherlands.

Brus, D.J. and de Gruijter, J.J., 1997: Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil. Geoderma 80: 1-59.

Bui, E.N., 2004: Soil survey as a knowledge system. Geoderma, 120: 17 – 26.

Bui, E.N., Loughhead, A., and Corner, R., 1999: Extracting soil-landscape rules from previous soil surveys. Australian J. of Soil Research, 37: 495-508.

Bui, E.N. and Moran, C.J., 2003: A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darlin basin of Australia. Geoderma, 111: 21-44.

Burrough, P.A. and McDonnell, R.A., 1998: Principles of Geographical Information System. 2nd ed., Oxford University Press, 356 pp.

Carre, F., McBratney, A.B., and Minasny, B., 2007: Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma, 141: 1- 14.

Chen, C., Liaw, A., and Breiman, L., 2004: Using Random Forest to Learn Imbalanced Data. Technical Report, 666. Department of Statistics, University of California, Berkely.

Cochran, W.G., 1977: Sampling Techniques, 3rd ed., John Wiley & Sons, 428 pp.

Debeir, O., Latinne, P., and van den Steen, I., 2001: Remote sensing classification of spectral, spatial and contextual data using multiple classifier systems. - Proccedings 8th ECS Image Analysis, Bordeaux: 584-589.

Domburg, P., de Gruijter, J.J., and van Beek, P., 1997: Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. Geoderma, 75: 183-201.

Ensor, K.B. and Glynn, P.W., 1997: Stochastic Optimization via Grid Search. IN: Mathematics of Stochastic Manufacturing Systems, American Mathematical Society: 399 pp.

Friedl, M.A. and Brodley, C.E., 1997: Decision tree classification of land cover from remotely sensed data, Remote Sensing Environment, 61: 399-409.

Friedman, J. H., 1991: Multivariate adaptive regression splines (with discussion). Annals of Statistics, 19/1: 1-82.

Geissen, V., Kampichler, C., López-de Llergo-Juárez, J.J., and Galindo-Acántara, A., 2007: Superficial and subterranean soil erosion in Tabasco, tropical Mexico: Development of a decision tree modeling approach. Geoderma, 139: 277 - 287.

Geng, W., Cosman, P., Berry, C.C., Feng, Z., and Schafer, W.R., 2004: Automatic Tracking, Feature Extraction and Classification of C. elegans Phenotypes. - IEEE Transactions on Biomedical Engineering, 10/51: 1811-1820.

Giasson, E., van Es, C., van Wambeke, A., and Bryant, R.B., 2000: Assessing the economic value of soil information using decision analysis techniques. - Soil Science 165/12: 971-978.

Gilardi, N. and Bengio, S., 2001: Local Machine Learning Models for Spatial Data Analysis. Journal of Geographic Information and Decision Analysis, 4/1: 11-28.

Gómez-Chova, L., Calpe, J., Soria, E., Camps-Valls, G., Martin, J.D., and Moreno, J., 2003: Cart-based Feature Selection of hyperspectral images for crop cover classification. - IEEE International Conference on Image Processing, 3: 589-592.

Gourieroux, Ch. and Monfort, A., 1995: Statsitcs and Econometric Models: General Concepts, Estimation, Prediction, and Algorithms. Cambridge University Press, Cambridge, 504 pp.

Grinand, C., Arrouays, D., Laroche, B., and Martin, M.P., 2008: Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma, 143: 180-190.

Gu, B., Hu, F., and Liu, H., 2001: Sampling: Knowing whole from its part. In Liu, H., and Motoda, H., 2001 (eds): Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, Boston, 448 pp.

Hansen, M., Dubayah, R., and DeFries, R., 1996. Classification trees: An alternative to traditional land cover classifiers. International Journal of Remote Sensing, 17, 1075–1081.

Hastie, T., Tibshirani, R., and Friedman, J.H., 2001: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 533 pp.

He, P., Fang, K.T., and Xu, C.-J., 2003: The Classification Tree Combined with SIR and Ist Application to Classification of Mass Spectra. Journal of Data Science, 1: 425-445.

Hole, F.D., 1978: An approach to landscape analysis with emphasis on soils. Geoderma, 21/1: 1-23.

Ishioka, T., 2003: Evaluation of Criteria for Information Retrieval. IEEE/WIC International Conference on Web Intelligence WI 2003, Sponsored by IEEE Computer Society and Web Intelligence Consortium, Halifax, Canada: 425-431.

Jenny, H., 1941: Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York, 281 pp.

Jin, Y., 2006: Multi-Objective Machine Learning. Springer Verlag, 660 pp.

John, G. H., Kohavi, R., and Pfleger, K., 1994: Irrelevant features and the subset selection problem. Proceedings of the 11th International Conference on Machine Learning, 121-129.

Kim, H. and Loh, W.-Y., 2001: Classification trees with unbiased multiway splits. Journal of the American Statistical Association, vol. 96, pp. 589-604.

Kohonen, T., Hynninen, J., Kangas, J., and Laaksonen, J., 1995: The Learning Vector Quantization Program Package; Version 3.1., Online Puplication: http://www.cis.hut.fi/research.

Lagacherie, P. and Holmes, S., 1997: Addressing geographical data errors in a classification tree for soil unit prediction. Int. J. Geographical Inf. Sci. 11, 183–198.

Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N., and Barthes, J.P., 2001: Mapping of reference area representativity using a mathematical soilscape distance. Geoderma 101: 105-118.

Liu, H. and Motoda, H., 1998: Feature Selection for knowledge discovery and Data Mining. Kluwer Academic Publishers, Boston, 214 pp.

Liu, H. and Motoda, H., 2001: Data Reduction via Instance Selection. In Liu, H. and Motoda, H., 2001: Instance Selection and Construction for Data mining. Kluwer Academic Publishers,

Boston, 448 pp.

Liu, H. and Motoda, H., 2002: On Issues of Instance Selection. Data Mining and Knowledge Discovery, Kluwer Academic Publishers, 6: 115-130.

Loh, W.Y. and Shih, Y.S., 1997: Split Selection Methods for Classification Trees. Statistica Sinica, 7: 815-840.

Loh, W.Y. and Vanichsetakul, N., 1988: Tree-structured classification via generalized discriminant analysis. Journal of the American Statistical Association, 83: 715-728.

Manning, C.D. and Schütze, H., 1999: Foundations of Statistical Natural Language Processing. MIT Press Cambridge (USA) und London, 620 pp.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003: On digital soil mapping. Geoderma 117: 3-52.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., and Shatar, T.M., 2000: An overview of pedometrics techniques for use in soil survey. Geoderma 97: 293-327.

McKenzie, N.J. and Ryan, P.J., 1999: Spatial prediction of soil properties using environmental correlation. Geoderma 89: 67-94.

Mitchell, T. M., 1997: Machine Learning. Singapore, 414 pp.

Mitchie, D., Spiegelhalter, D.J., and Taylor, C.C., 1994: Machine Learning, Neural and Statistical Classification. Ellis Horwood, New York, 298 pp.

Moran, J.C. and Bui, E.N., 2002: Spatial data mining for enhanced soil map modeling. Int. J. Geographical Information Science, 16/6: 533-549.

Munoz, J. and Felicisimo, A.M., 2004: Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science, 15: 285-292.

Pal, N.R. and Jain, L., 2005: Advanced Techniques in Knowledge Discovery and Data Mining. Springer Verlag, London, 264 pp.

Qi, F., 2004: Knowledge discovery from area-class resource maps: Data preprocessing for noise reduction. Transactions in GIS, 8/3: 297-308.

Qu, P. and Loh, W.Y., 1992: Application of Box-Cox transformations to discrimination for the two-class problem, Communications in Statistics (Theory and Methods) 21: 2757-2774.

Raghavan, V., Bollmann, P., and Jung, G. S., 1989: A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Inf. Syst., 7: 205-229.

Schafer, J., 1997: Analysis of incomplete multivariate data. Chapman & Hall, London. 430pp.

Scull, P., Franklin, J., and Chadwick, O.A., 2005: The application of classification tree analysis to soil type prediction in a dessert landscape. Ecological Modelling, 181: 1 - 15.

Shrestha, D.P., Zinck, J.A., and Van Ranst, E., 2004: Modelling land degradation in the Nepalese Himalaya. Catena: 135-156.

Van Rijsbergen, C.J., 1979: Information Retrieval. Butterworths, 153 pp.

Wai, L., Keung, C.-K., and Ling, C.X., 2001: Learning via prototype generation and filtering. In Liu, H. and Motoda, H., 2001 (eds): Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, Boston, 448 pp.

Wilson, D.L., 1972: Asymptotic properties of nearest neighbour rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics 2: 408-420.

IUSS Working Group, 2006: World reference base for soil resources 2006. 2nd ed., World Soil Resources Reports No. 103, FAO, Rome.

Zhang, J., Guo, D., and Wan, Q., 1999: Geospatial Data Mining and Knowledge Discovery using Decision Tree Algorithm - A Case Study of Soil Data Set of the Yellow River Delta. - Geoinformatics and Socioinformatics, Proceedings of Geoinformatics'99 Conference, Ann Arbor, Michigan: 1-8.

Zhou, B., Zhang, X., and Wang, R., 2004: Automated soil resources mapping based on decision tree and Bayesian predictive modeling. Journal of Zhejiang University Science, 5: 782-795.

Zhu, A.X., 2000: Mapping soil landscape as spatial continua: The neural network approach. - Water Resources Research 36/3: 663-677.

Manuscript 4

An approach to removing uncertainties in nominal environmental covariates and soil class maps

Hartemink, McBratney & Mendoca-Santos (2008): Digital Soil Mapping with limited data. Springer.

Thorsten Behrens, Karsten Schmidt, Thomas Scholten

Institute of Geography, Chair of Physical Geography, Eberhard Karls University Tübingen, Rümelinstraße 19-23, D-72074, Tübingen, Germany

Abstract

In this chapter we present an automated approach to correct the delineation of nominal soil and environmental datasets based on auxiliary metric attributes, aiming to enhance positional accuracy. The detection of uncertainties is based on different spatial and non-spatial approaches. The methodological framework mainly consists of nearest neighbour approaches and comprises supervised feature selection, different ensemble classification techniques, as well as spatial and non-spatial smoothing and generalization approaches. The method is described and applied to an artificial dataset as well as a 1:50 000 German soil map and a 1:1 000 000 geological map of the Republic of Niger.

1 Introduction

In many situations of applied digital soil mapping we have to handle spatial datasets of varying provenance, age, scale, resolution, mapping scheme, and aggregation level resulting in different sources of errors (Robinson et al., 1984; Lagacherie and Holmes, 1997; Heuvelink, 1998; Bishop et al., 2006). In predictive data mining approaches (Behrens et al., 2005; Behrens and Scholten, 2006a) existing soil data is extrapolated on the basis of auxiliary environmental datasets (McBratney et al., 2003). Hence, the prediction accuracy can be weak i) if the soil data and or ii) the auxiliary datasets contain errors. For example when using a small-scale geological maps (> 1:100 000) as predictor datasets for medium or large scale digital soil maps (< 1:50 000) the delineation is propagated through the analysis and can be found in the prediction results, assuming there is a significant relation. Thus, in general, maps of smaller scales should not be used to compile maps of larger scales (Robinson et al., 1984). As “in practice even the best-drawn maps are not perfect” (Burrough and McDonnell, 1998) positional inaccuracies can be

found in most classically surveyed soil maps for many well known reasons. In this case, the soil map as the training dataset contains noise - which again weakens prediction accuracy (Brighton and Mellish, 2002). Hence, it is important to provide solutions to automatically correct existing datasets. One step towards better predictions - or in some cases to allow predictions at all, when the data is sparse and at small scale - is to provide techniques that correct the boundaries of nominal datasets on the basis of auxiliary datasets of higher resolutions and/or scales. Demonstrated on different artifical and real datasets this study presents automated approaches to adjust the boundaries of nominal datasets based on terrain attributes.

2 Rationale

The correction of the delineation in nominal datasets as introduced in this paper is based on two major steps: first, the detection and removal of inaccuracies and second, the prediction of new class values for all incorrect pixels. The detection of positional inaccuracies can be achieved in an unsupervised or a supervised fashion, based on simple band removal approaches or on outliers in terrain attributes found within each class-area. Concerning the outlier based approach digital terrain analysis plays a crucial role. Hence a large library of terrain attributes (Behrens, 2003; Behrens et al., 2005) is used. The prediction of a new class value for an uncertain pixel is based on a spatial and non-spatial nearest neighbour data mining framework, comprising feature selection, ensemble classifications as well as spatial and non spatial smoothing and generalization approaches to provide stabe and reliable results.

The whole approach is applied iteratively, as the position of the boundaries and thus the outliers changes after each run. The system is stopped if no further significant or plausible changes occur.

3 Test sites

3.1 Artificial Datasets

An artificial DEM of a hemisphere (radius = 40 pixels) set on top of a plain surface (100 by 100 pixels) was used as a test bed for the framework developed here. The corresponding artificial nominal environmental dataset consists of two mapping units: first, an inner, irregular shaped ellipsoid which overlaps large parts of the hemisphere and minor parts outside the hemisphere and second a surrounding mapping unit mainly covering the plain surface surrounding the hemisphere (Fig. 1). The aim is to correct the ellipsoid mapping unit in such way that it covers the hemisphere completely and is removed from the surrounding plain surface.

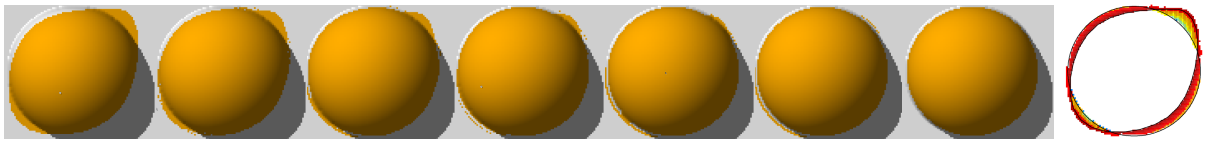


Fig. 1. Iterative correction of positional inaccuracies of an ellipsoid mapping unit to an artificial hemisphere DEM. The left image is the original ellipsoid followed by images showing the results after 2, 4, 6, 8, 10 and 12 iterations. The image in the left shows the location of the corrected pixels.

3.2 1:50.000 soil map of Central Hesse, Germany

The 1:50.000 soil map of central Hesse, Germany comprises the Vogelsberg, Europe's biggest shield volcano with a relief of 170 m asl. to 750m asl.. The soilscape is mainly characterized by Cambisols, often influenced by loess components (HLUG, 2002). The soil map was rasterized to a resolution of 20 meters.

3.3 1:1.000.000 geological map of the Republic of Niger

To provide an example for countries with sparse datasets, we analyzed the 1:1.000.000 geological map of the Republic of Niger in Western Africa. The map is based on the work of Greiert (1961) and disseminated digitally as part of the "Atlas of Natural and Agronomic Resources of Niger and Benin" (Herrmann et al., 1999). Shuttle Radar Topography Mission (SRTM) data with a resolution of 90 m was used to derive terrain attributes (cf. Chapter 4.4.1).

4 Methods

The technical key steps of the proposed methodology are briefly described in the following:
Beyond the correction of uncertainties the major goal of the methodological framework applied is to provide stable results. Thus in the first step after digital terrain analysis (cf. Chapter 4.4.1) and dataset creation, the data are analyzed to remove noisy and irrelevant features (attributes) (cf. Chapter 4.4.2). After reducing dimensionality, the core algorithms, that is, removing uncertain pixels form the dataset, are applied (cf. Chapter 4.4.3). Afterwards, to speed up computation time we use a stratified random sampling (cf. Chapter 4.4.5.1) over the resulting dataset followed by Wilson editing to remove noise (cf. Chapter 4.4.5.2). To allocate each uncertain pixel we apply a simple kNN-classifier (cf. Chapter 4.4.5.3) to assign the most probable class out of the spatially adjacent neighbours (cf. Chapter 4.4.4). Finally to keep the system stable and to avoid blurry borders local spatial noise removing is applied (cf. Chapter 4.4.6). Additionally the methods described above are embedded in an ensemble prediction approach (cf. Chapter 4.4.7) - again to provide accurate results.

4.1 Digital terrain analysis

As there is a strong dependence between terrain attributes and soil properties (McBratney et al., 2001; Behrens et al., 2005), a large pool of continuous geomorphometric terrain attributes is used in digital soil mapping. Based on a terrain-analysis framework (Behrens, 2003), the following 25 terrain attributes were calculated: flow accumulation, relative hillslope position, elevation above channel, distance to channel, average slope, steepest slope, aspect, profile curvature, planform curvature, mean curvature, maximum curvature, minimum curvature, relative profile curvature, relative planform curvature, topographic roughness, relative richness, waxing and waning slopes, solar insolation, compound topographic index, USLE LS-factor, landform evolution, relative mass balance, stream power index, surface area ratio, and surface volume. For details see Behrens (2003), Behrens et al. (2005), and Behrens and Scholten (2006b).

SRTM data was used to remove uncertainties in the geological map of the Republic of Niger. Additional terrain attributes had to be calculated based on Monte Carlo simulations to derive more natural spatial flow patterns and geomorphometric positions (i.e. flow accumulation, relative hillslope position, and elevation above channel). This was essential for boundary adjustments, due to the error component of the SRTM data which is relatively large in this area (visual interpretation), the scale of the geological map, and the width of the valleys. A comparison of a standard multiple-flow algorithm to calculate contributing area (Dietrich and Montgommery, 1998) and a Monte Carlo approach based on the D8 single-flow algorithm (Jenson and Dominique, 1988) is shown in Figure 2 [a, b]. It can be seen that the spatial distribution of the resulting flow accumulations based on the different approaches differs especially in the valley bottoms, where the Monte Carlo based approach produces much more plausible results and models the valley bottom according to the draped hillshade.

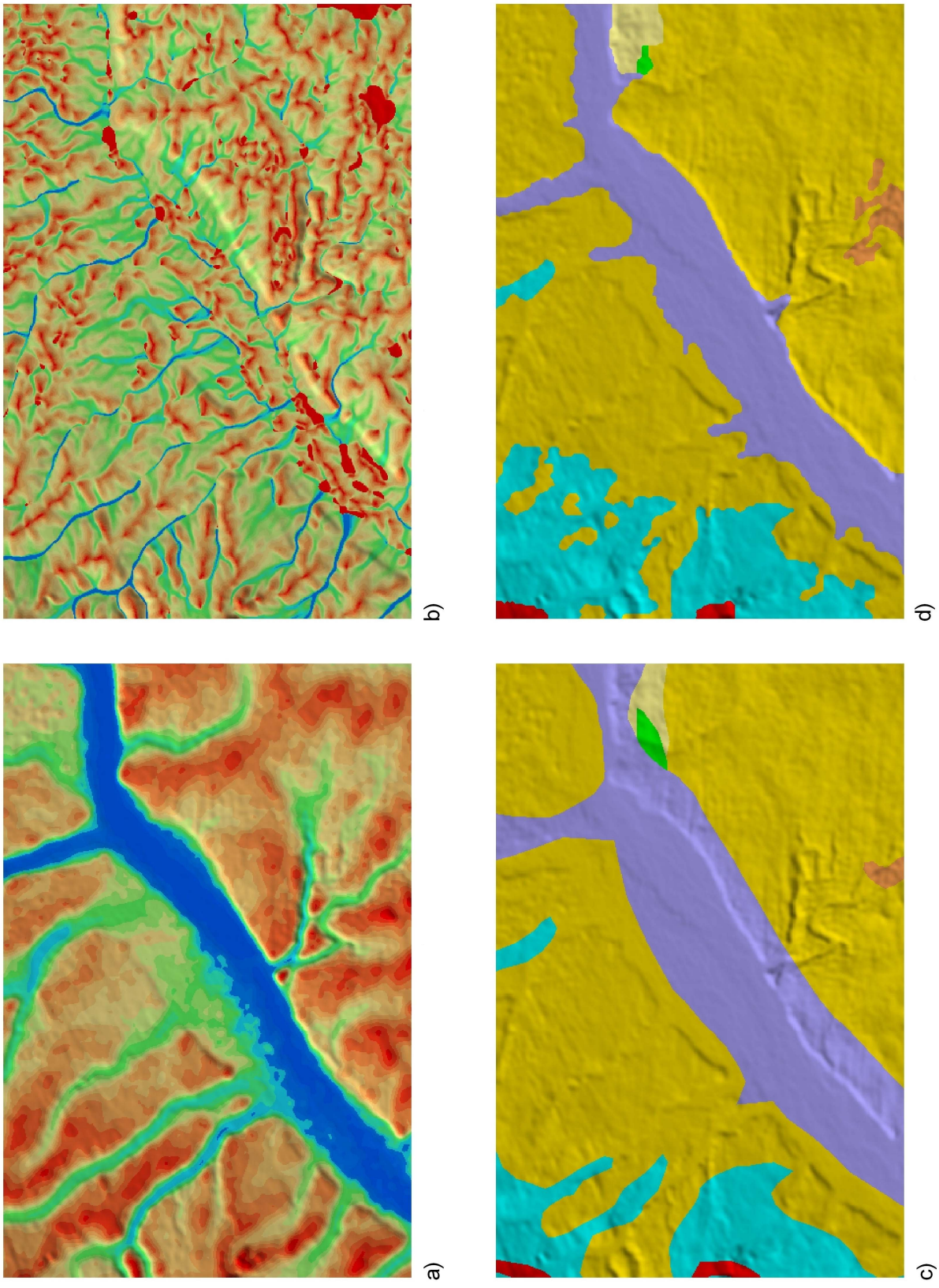


Fig. 2. Comparison of flow-accumulation based on a Monte Carlo simulation using a single flow approach [a] and a multiple-flow approach [b] for a section of the geological map of the Republic of Niger [c = original, d = corrected].

4.2 Feature selection

K-nearest neighbour classifiers as instance-based learners are sensitive to correlated as well as irrelevant and noisy features. Thus, feature selection techniques need to be applied to achieve accurate predictions. The feature selection algorithm used in this study is the well known Relief-F approach (Kira and Rendell, 1992; Kononenko, 1994; Liu and Motoda, 1998).

For every class combination and every randomly selected instance (i.e. vector containing terrain attributes) in a dataset the difference between the feature values of the nearest hit (i.e. the shortest instance to the same class) and the nearest miss (i.e. the shortest instance to the neighbouring class) are calculated and summed up over all selected instances in a weight vector. Thus each feature has a weight indicating its potential to differentiate between the classes in a dataset. In this study we used the mean weight as the lower limit to remove features and 50 randomly selected instances per class.

4.3 Removing Uncertainties

4.3.1 Band removal

As the probability of noise is generally higher at the polygon boundaries than within a polygon, a simple spatial denoising approach is to remove all pixels at the class boundaries. This idea is based on the concept of “error bands” as introduced by Perkal (1966). As the width of the band can not easily be predicted and is irregular in most cases we use a buffer width of one pixel inside each class.

4.3.2 Outlier detection

Based on the terrain attributes derived (cf. Chapter 4.4.1) we developed a non-spatial denoising approach which is the initial idea behind this study. Therefore, each class-area of the nominal dataset to be corrected is analyzed separately in terms of outliers within the frequency distribution of each terrain attribute. If a threshold is reached, that is if the majority of all terrain attributes is outside twice the standard deviation, the corresponding pixel is marked as uncertain. Thus, this process is data driven and in contrast to band denoising it is not fixed to the boundary between two polygons.

4.3.3 Spatial neighborhood search

To determine the most likely soil class for each uncertain pixel a local spatial neighbourhood search is applied. The advantage of a local search procedure is that the universe of potential

classes to be assigned to an uncertain pixel is reduced resulting in predictions that are generally more accurate and stable.

The search for adjacent soil classes is based on the moving window technique. If a pixel contains no soil class information, its neighborhood is analyzed initially on the basis of a three-by-three pixel neighbourhood. If no adjacent soil class is found within this kernel, the neighborhood size is automatically enlarged until at least two adjacent soil classes are found.

4.4 Instance selection and classification

4.4.1 Random sub-sampling

Instance selection, or sub-sampling (Liu and Motoda, 2001, Manuscript 3) aims to remove redundant information from datasets as well as to speed up learning and/or prediction time while preserving prediction accuracy. This becomes important for large datasets with thousands of training samples and for computationally expensive algorithms like *k-nearest neighbour*.

In this study we use stratified random sampling. Kohonen et al. (1995) recommend a disproportional approach for supervised classification applications, where an equal amount of observations or instances is selected for each class, even when the *a-priori* probabilities differ strongly.

The sample size for each class in this study is 50. As this process is embedded in the ensemble prediction approach random sampling is the subagging (Breimann, 1996, Bühlmann and Yu, 2002) part of the procedure.

4.4.2 Dataset editing

Wilson editing (Wilson, 1972) is a competitive supervised denoising technique (Zeidat et al., 2005) with the goal of obtaining more accurate classifiers. This is achieved by removing all vectors in a dataset that have been misclassified by a k-nearest neighbour classifier, leading to smoother class boundaries in the feature space and better subsequent classification results using a k-nearest neighbour classifier. In our case this is done separately for every class combination found within the neighborhood search to get optimized and spatially dependent classification results, which is an advantage in heterogeneous soilscapes.

4.4.3 Supervised classification

The k-nearest neighbour classifier (Fix and Hodges, 1951) labels an unknown instance with the class label of the majority of its k-nearest neighbours in terms of Euclidean distance in feature

space. We use a three-nearest neighbour classifier in this study.

4.5 Generalization using spatial noise removal

To avoid blurry borders and small isolated areas a spatial noise removal is applied after each iteration. The noise removal approach replaces all areas comprising less than five pixels followed by a three-by-three pixel majority filter. An optimized size for areas to be removed can not be determined a-priori and has to be tested iteratively. We suggest a default value of less than 5 pixels (which is less than the half amount of pixels in a local 3*3 three-by-three neighbourhood), as it produces only a weak smoothing effect. As this process avoids fuzzy transition zones due to smoothing it is comparable to the soil scientist's approach of generalization when mapping soil classes.

4.6 Ensemble prediction

Ensemble approaches, i.e. combinations of multiple predictions based on changes in the training dataset, are very popular and powerful, as they increase prediction accuracy (Breiman, 1996). For k-nearest neighbour classifiers, feature subset selection approaches are recommended by a number of authors (e.g., Bay, 1999; Akkus and Güvenir, 1996) and are competitive with boosted (Freund and Schapire, 1996) decision trees (Bay, 1999). The application of instance-based ensembles like bagging or subagging (Breimann, 1996; Andonova et al., 2002, Bühlmann and Yu, 2002) is reported to work on small random samples (Alpaydin, 1997; Hamamoto et al., 1997). Additionally ensemble approaches are more robust against irrelevant and correlated features (Bay, 1999; Skurichina and Duin, 2001). In this Chapter we apply a combination of both random feature subsets (Ho, 1998) as well as small sample subagging (cf Chapter 4.4.5.1), which is comparable to the decision forest approach (Ho, 2001) for decision trees.

5 Results and Discussion

The aim of applying the proposed approach on an artificial dataset was to visualize the results for an easy-to-interpret example. As shown in Fig. 1 the method works as expected. Based on the band removal approach 12 iterations were needed to fit the nominal ellipsoid to the hemisphere (circle) based on 5 terrain attributes (slope, compound topographic index, mean, profile and horizontal curvatures). Using the outlier detection and a low threshold for outlier removal, only one iteration is needed to achieve the same results. Yet in this case band removal offers the opportunity to analyze the correction process and thus the location of spatial uncertainties over the iterations.

The correction of the 1:50.000 soil map of Central Hesse, Germany is based on the outlier detection approach (cf. Chapter 4.4.3.2). A first field survey revealed four iterations to be sufficient to correct the soil map. At first sight, only minor changes can be found between the original soil map (Fig. 3) and the corrected soil map (Fig. 4).

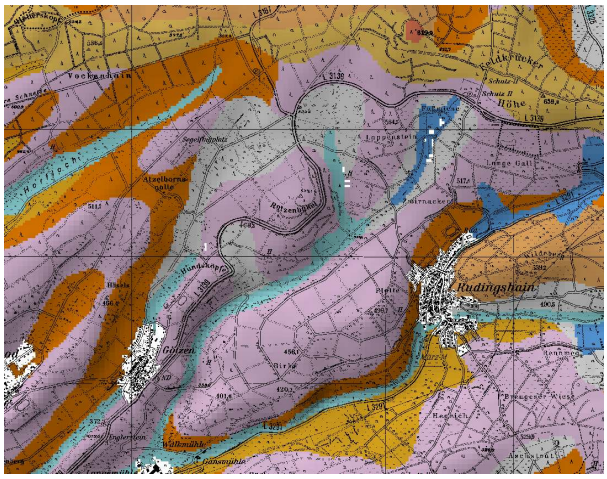


Fig. 3. Section of the original 1:50 000 soil map (Central Hesse, Germany) draped over a DEM.

Yet, some boundaries show differences of 80 m – resulting from 4 iterations on a 20 m grid dataset. Hence, using a finer resolution more iterations would be required. Depending on the resolution and the scale of the input datasets different resolutions need to be tested to achieve optimized results.

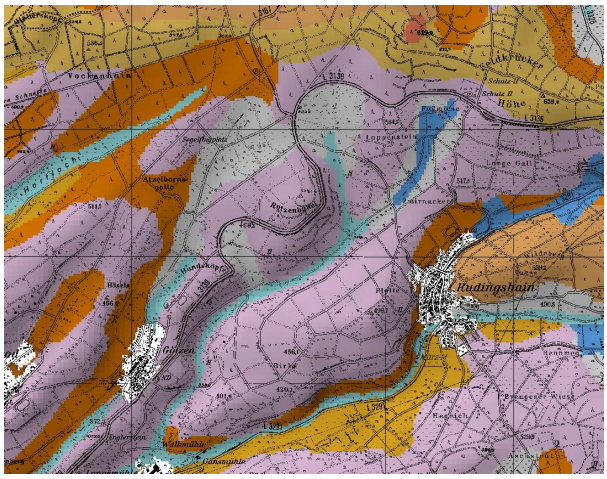


Fig. 4. Corrected 1:50 000 soil map (Central Hesse, Germany) draped over a DEM.

The section of the 1:1 000.000 geological map of the Republic of Niger as shown in Fig. 2 [c] demonstrates the problem of small-scale nominal maps in relation to more precise, in terms of positional accuracy, data like SRTM. For the mapped floodplain in the centre of the section, 33 % of the area is located outside the floodplain as shown by the SRTM DEM. Thus, automated-data-mining-based boundary adjustments become critical due to large amounts of noise. Yet, 12 iterations based on the outlier approach were sufficient to correct the map (Fig. 2 [c,d]).

Attempts with standard terrain attributes only, were not successful, whereas a combination with the Monte Carlo based approaches returned reasonable results. In the case of some of the other mapping units which do not show a strong relation to relief, we recommend not using the adjusted polygons in the final map.

Based on the three datasets tested during the development of the approach presented the technique works as expected. Yet, it is not possible to make assumptions about the quality of the boundary adjustment a priori as it depends on the errors and interactions of the datasets used. Generally, three approaches to test the accuracy of the method seem reasonable: i) direct field validations, ii) indirect comparisons between digital soil maps based on the original and on the corrected datasets, and iii) expert interpretations based on map overlays using hillshades or 3D visualisations. Concerning the 1:50.000 soil map of Hesse all 3 approaches are currently tested and compared for different landscapes.

Concerning different methodological techniques (iterations, removal techniques) the general tendency is that more iterations are needed if the uncertainties are high or the resolution of the terrain attributes is relatively fine compared to the scale of the map to correct. Further research is needed on criteria when to stop the iterative process to provide fully automated applications. Band removal needs more iterations than outlier based removal. Additionally the risk to change boundaries not related to relief is higher. Finally, for geomorphologic settings like wide valleys as found in the geological map of Niger, plausible results can only be achieved when carefully selected special terrain attributes are used in the correction approach.

6 Conclusions

The approach introduced in this study helps to enhance positional accuracy of nominal soil and environmental datasets. This is important:

- in terms of error propagation if datasets of different scales and resolutions are used together in *scorpan* (McBratney et al., 2003) predictions,
- to produce better prediction results due to an enhancement of the soil map to predict and / or the environmental covariates used for prediction,
- and in terms of quality control of conventionally produced soil maps.

The second step of the procedure - the prediction of the most probable class - can also be used as a post-processing tool after predicting single soil units which often results in overlapping areas and gaps (Behrens et al., 2005).

Future research is needed to find optimized settings for the different model parameters like the feature selection threshold, the removal approach, the sub-sample size, the number of

neighbours used for classification, the settings for spatial smoothing, and the iterations. Yet, as using the default values returns promising and stable results the major model parameters are the number of iterations and the algorithm to remove noise. The principal advantage of the outlier-based removal compared to the error-band removal is that pixels at class boundaries that are not related to the environmental covariates used, are not per se regarded as uncertain, which preserves these boundaries in their original shape. Hence, outlier-based removal is recommended.

The method, proposed and described here for the first time, may become an important pre- and post-processing tool in digital soil mapping. In countries with sparse and coarse soil information it might help to refine maps of soil and its environmental covariates.

Acknowledgements

The authors would like to gratefully acknowledge the critical and helpful comments and suggestions received from Alex McBratney on this manuscript. Partial funding for this research was provided by the Collaborative Research Centre 299 of the German Research Foundation. We would like to thank the Hessian State Office for Environment and Geology (HLUG) for providing data.

References

Andonova, S., Elisseeff, A., Evgeniou, T., Pontil, M., 2002. A Simple Algorithm for Learning Stable Machines. In Frank van Harmelen (Ed.): Proceedings of the 15th Eureopean Conference on Artificial Intelligence. IOS Press, Amsterddam. pp. 513-517.

Akkus, A., Güvenir, H.A., 1996. K nearest neighbor vlassification on feature projections.. In: Lorenza Saitta (Ed.), Machine Learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, San Francisco, CA. pp. 12-19.

Alpaydin, E., 1997. Voting over multiple condensed nearest neighbors. Artificial Intelligence Review 11, pp. 115–132.

Bay, S.D., 1999. Nearest neighbor classification from multiple feature subsets. Intelligent Data Analysis 3, pp. 191-209.

Behrens, T., 2003. Digitale Reliefanalyse als Basis von Boden-Landschafts-Modellen – am Beispiel der Modellierung periglaziärer Lagen. Boden und Landschaft 42, 189p. Giessen.

Behrens, T., Förster, H. Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. and Soil Sci. 168, pp. 21-33.

Behrens, T., Scholten, T., 2006a. Digital Soil Mapping in Germany – a review. J. Plant Nutr. and Soil Sci. 169, pp. 434 - 443**.**

Behrens, T., Scholten, T., 2006b. Chapter 25. A comparison of data-mining techniques in predictive soil mapping. In Lagacherie, P. McBratney, A.B., Voltz, M. (Eds): Digital Soil mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam. pp. 353-364.

Bishop, T.F.A., Minasny, B., McBratney, A.B., 2006. Uncertainty analysis for soil-terrain models. International Journal of Geographical Information Science 20, pp. 117-134.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, pp. 123-140.

Breiman, L., 2001. Random forests. Machine Learning 45, pp. 5-32.

Brighton, H., Mellish, C., 2002. Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery 6, pp. 153-172.

Bühlmann, P., Yu, B., 2002. Analyzing bagging. Annals of Statistics 30, pp. 927-961.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford University Press, New York.

Dietrich, W.E., Montgomery, D.R., 1998. SHALSTAB: a digital terrain model for mapping shallow landslide potential, NCASI (National Council of the Paper Industry for Air and Stream

Improvement), Corvallis OR. Technical Report, 29p.

Fix, E., Hodges, J.L., 1951. Discriminatory analysis - Nonparametric discrimination: Consistency properties. Technical Report Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolf Field, Texas.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm, In: Lorenza Saitta (Ed.), Machine Learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, San Francisco, CA. pp. 148-156.

Greigert, J., 1961. République du Niger. Carte géologique de reconnaissance du Bassin des Iullemeden 1:1 Mio. BRGM, Niamey, Niger.

Hamamoto, Y., Uchimura, S., Tomita, S., 1997. A bootstrap technique for nearest neighbor classifier design. IEEE Trans. Pattern Analysis and Machine Intelligence 19, pp. 73-79.

Herrmann, L., Vennemann, K., Stahr, K., von Oppen, M., 1999. Atlas of Natural and Agronomic Resources of Niger and Benin.

http://www.uni-hohenheim.de/~atlas308/startpages/page2/english/content/title_en.htm

Heuvelink, G.B.M., 1998, Error Propagation in Environmental Modelling with GIS. Taylor & Francis, London. 144p.

HLUG (Hessische Landesamt für Umwelt und Geologie), 2002. Erläuterungen zur Bodenkarte von Hessen 1:50000. HLUG, 575p.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, pp. 832-844.

Jenson, S.K., Domingue, J.O., 1988. Extracting topographic structure from digital elevation data for Geographic Information System analysis, Photogrammetric Engineering and Remote Sensing 54, pp. 1593-1600.

Kira, K., Rendell, L., 1992. A practical approach to feature selection. In Derek H. Sleeman, Peter Edwards (Eds.): Proceedings of the Ninth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA., pp. 249-256.

Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J., 1995. The Learning Vector Quantization Program Package; Version 3.1. http://www.cis.hut.fi/research/lvq_pak/.

Kononenko, I., 1994. Estimating attributes: Analysis and extensions of RELIEF. In De Raedt, L, Bergadano, F. (Eds.): European Conference on Machine Learning. Springer, Heidelberg. pp. 171–182.

Lagacherie, P., Holmes, S., 1997. Addressing geographical data errors in a classification tree soil

unit prediction. International Journal of Geographical Information Science 11. pp.183-198.

Liu, H., Motoda, H., 1998. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Norwell, MA. 244p.

Liu, H., Motoda, H., 2001. Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, 448p.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117. pp. 3-52.

Perkal, J., 1966. An Attempt at Objective Generalization. Michigan Inter-University Community of Mathematical Geographers. Discussion Paper 10, Univ. of Michigan.

Robinson, A.H., Sale, R.D., Morrison, J.L. and Muehrcke, P.C., 1984, Elements of Cartography, 5th Edition, John Wiley & Sons, New York, NY.

Skurichina, M. and Duin, R.P.W., 2001. Bagging and the random subspace method for redundant feature spaces. In J. Kittler, F. Roli (Eds.): Multiple Classifier Systems, Proceedings Second International Workshop MCS 2001 (Cambridge, UK, July), Lecture Notes in Computer Science, vol. 2096, Springer Verlag, Berlin pp. 1-10.

Wilson, D.L., 1972. Asymptotic properties of nearest neighbour rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics 2, pp. 408-420.

Zeidat, N., Wang, S., Eick, C., 2005: Dataset Editing Techniques: A Comparative Study, in revision.

Curriculum Vitae

Dipl. Geogr. Karsten Schmidt

Eberhard-Karls-Universität Tübingen
Institut für Geographie
Rümelinstr. 19-23, 72070 Tübingen
Tel.: +49 (0) 7071 - 29 - 77523
Fax: +49 (0) 7071 - 29 - 5391
E-mail: Karsten.Schmidt@uni-tuebingen.de

Geburtsdatum:	28.02.1978
Geburtsort:	Burg
Nationalität:	deutsch

Ausbildung

01.2009	Promotionsthema: „ Repräsentanz und Data Mining – Konzepte und Methoden der digitalen bodenkundlichen Kartierung“
01.2006 – 01.2009	Promotionsstudent am Lehrstuhl für Physische Geographie, Institut für Geographie und Bodenkunde, Eberhard-Karls Universität Tübingen im Sonderforschungsbereich (SFB) 299 „Landnutzungskonzepte für periphere Regionen“ der Deutschen Forschungsgemeinschaft (DFG) an der Justus-Liebig Universität Gießen
03.2007	Sino-German Summer School: Concepts and Algorithms in Geoecosystem Modelling (CAGEM), Yichang, China
12.2005	Abschluss Diplomarbeit mit dem Thema: „Vergleichende Analyse überwachter Klassifikationsverfahren in der Digitalen Bodenkartierung – Ein Methodenvergleich am Beispiel des mittleren Pfälzer Waldes“
10.1998 – 12.2005	Studium der Geographie mit den Hauptfächern Physische Geographie und Geoinformatik an der Friedrich-Schiller Universität Jena [Dipl. Geogr.]
06.1996	Abitur, Laurentinum Gymnasium Loburg, Sachsen-Anhalt

Beruflicher Werdegang

seit 11.2008	Wissenschaftlicher Angestellter im EU-Projekt iSoil „Interaction between soil related sciences - Linking geophysics, soil science, and digital soil mapping“ an der Eberhard-Karls Universität Tübingen
01.2006-10.2008	Wissenschaftlicher Angestellter im Sonderforschungsbereich 299 „Landnutzungskonzepte für periphere Regionen“ der Deutschen Forschungsgemeinschaft (DFG) an der Justus-Liebig Universität Gießen
09.2005 – 01.2007	Wissenschaftliche Projekte zum Thema Digitale Bodenkartierng mit verschiedenen Kooperationspartnern, wie dem Wasserforschungsinstitut der ETH-Zürich (EAWAG), dem Institut für Pflanzenernährung und Bodenkunde der Universität Halle, der Thüringer Landesanstalt für Umwelt und Geologie, dem Landesamt für Geologie und Bergwesen Sachsen-Anhalt
10.2003 – 12.2005	Studentische Hilfskraft im Fachbereich Physische Geographie und Bodenkunde, Institut für Geographie, Friedrich-Schiller Universität Jena
01.2000 – 03.2002	Studentische Hilfskraft am Lehrstuhl für Geoinformatik, Institut für Geographie, Friedrich-Schiller Universität Jena

Danksagung

Für das Vertrauen in meine Arbeit, die konstruktiven und auch kritischen Gespräche zum Thema meiner Promotion und nicht zuletzt für die ausgezeichnete Unterstützung meiner Arbeit möchte ich mich ganz herzlich bei Herrn **Prof. Dr. Thomas Scholten** bedanken. Ich freue mich auf eine weiterhin fruchtbare Zusammenarbeit.

Weiterhin möchte ich mich bei Herrn **Prof. Dr. Volker Hochschild** bedanken, der schon in meiner Studienzeit in Vorlesungen und Seminaren mein Interesse im Fachbereich GIS gefördert und mitentwickelt hat. Ganz besonders freue ich mich, ihn als einen Gutachter meiner Dissertation gewonnen zu haben.

Die fruchtbare Zusammenarbeit im Sonderforschungsbereich 299 brachte eine intensive Kooperation mit dem Institut für Bodenkunde und Bodenerhaltung der Justus-Liebig Universität Gießen mit sich. Für die Möglichkeit und die Unterstützung dieser intensiven Zusammenarbeit, als auch für die freundliche Übernahme eines Gutachtens, möchte ich mich bei Herrn **Prof. Dr. Peter Felix-Henningsen** bedanken.

Ein besonderer Dank gilt Herrn **Dr. Thorsten Behrens**, seine unermüdliche Art, Themen konsequent und umfassend anzugehen und zu bearbeiten waren und sind mir stets Vorbild. Die unendlichen Gespräche über die verschiedenen Teilaspekte meiner Arbeit waren immer produktiv und führten zu einer kontinuierlichen Verbesserung. Ihn als Kollegen und Freund zu bezeichnen erfüllt mich mit Stolz.

Bei meinem Kollegen und Freund Herrn **Dipl. Geogr. Christian Albrecht** möchte ich mich herzlich für die offene und freundliche Art bedanken, die eine herausragende Zusammenarbeit im Rahmen des SFB 299 ermöglichte. Sein umfassendes Wissen in der bodenkundlichen Geländeaufnahme hat meine Arbeit gestärkt und viele neue technische Aspekte unterstützt.

Für die hilfreichen Anmerkungen in beruflichen Belangen und für sein offenes Ohr in Diskussionen und Gesprächen möchte ich mich bei meinem Kollegen und Freund **Dipl. Geogr. Ralf Gründling** bedanken.

Für die freundliche Unterstützung und die Bereitstellung der umfangreichen Datengrundlagen im Einzugsgebiet der Nidda und darüber hinaus möchte ich mich bei den Verantwortlichen im Hessischen Landesamt für Umwelt und Geologie bedanken.

Den vielen fleißigen Händen seitens der Studierenden des geographischen Instituts sei an dieser Stelle gedankt. Durch Exkursionen und Geländeaufenthalte im Untersuchungsgebiet konnte nicht nur der Wissensdurst unserer Studierenden gestillt, sondern auch das ein ums andere Mal meine Sichtweise der Dinge erweitert werden. Ganz besonders möchte ich mich bei Frau **Ann-Kathrin Schatz**, Herrn **Markus Becker** und Herrn **Jonas Daumann** bedanken, die mir zeigten,

das begeisterungswillige Studierende für den Fachbereich Bodenkunde und Pedometrie nicht ausgestorben sind.

Meinen Eltern sei für ihre immerwährende Unterstützung in allen Lebensbereichen ganz besonders gedankt. Bei schwierigen Entscheidungen waren sie stets an meiner Seite, haben mir den Rücken gestärkt und nie an mir gezweifelt. Allein dafür gebührt ihnen nicht nur Dank sondern Hochachtung.

Obwohl sie meine kleine Schwester ist, habe ich viel von ihr gelernt. Susanne auch bei dir möchte ich mich recht herzlich bedanken. Die vielen kleinen Dinge, die mich und diese Arbeit unterstützten sind zahlreich und so möchte ich mich besonders für das Korrekturlesen bedanken.

Nicht zuletzt geht ein ganz besonderer Dank an Magdalena. Ihre Liebe und Unterstützung hat mich in allen Lebenslagen stets gestärkt und aufgebaut. Ich bin überglücklich, dass ihre Antwort auf meine Frage „Ja“ war.

Wissenschaftliche Publikationen und wichtige wissenschaftliche Beiträge

Behrens, T.**, Schmidt, K.**, Zhu, A.-X., and Scholten, T. (2009). Topography revisited - The ConMap approach for terrain based digital soil mapping. EJS, (submitted).

Schmidt, K., Behrens, T. & Scholten, T. (2009). Generation of soilscapes by segmenting soil maps for digital soil sensing and mapping in homogeneous feature spaces. J. Plant Nutrition and Soil Science (accepted).

Behrens, T., Zhu, A.X., **Schmidt, K.,** Scholten, T. (2009). Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma (accepted).

Behrens, T., **Schmidt, K.,** Gerber, R., Albrecht, C., Felix-Henningsen, P., Scholten, T. (2009). Concepts for generating shortest representative transects – sampling approaches for linear operated proximal soil sensors. J. Geogr. Inf. Science (accepted).

Baumann, F., He, J.S., Kühn, P., **Schmidt, K**., Scholten, T. (2009): Pedogenesis, permafrost, and soil moisture as controlling factors for soil nitrogen and carbon contents across the Tibetan Plateau. Global Change Biology (accepted).

Schmidt, K., Behrens, T. & Scholten, T. (2008). Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. Geoderma, 146, 138-146.

Behrens, T., **Schmidt, K.,** Gerber, R., C. Albrecht, C., Felix-Henningsen, P., Scholten, T. (2008). Shortest representative transects for linear operated proximal soil sensing surveys. In: Viscarra-Rossel et al.: Proc. 1st Global Workshop on High Resolution Digital Soil Sensing and Mapping. Sydney, Australia.

Albrecht, Ch., **Schmidt, K.,** Gerber, R., Behrens, T., Felix-Henningsen, P. & Scholten, T. (2007). Georadaruntersuchungen repräsentativer Transekte im Einzugsgebiet der Nidda (Hessen). Mitteilgn. Dtsch. Bodenkundl. Gesellsch. 110(2): 423f.

Behrens, T., **Schmidt, K.** & Scholten, T. (2008). An approach to remove uncertainties in nominal environmental covariates and soil class maps. In: Hartemink, McBratney & Mendoca-Santos (2008): Digital Soil Mapping with limited data. Springer.

Schmidt, K., Behrens, T. & Scholten, T. (2007). Landschaftssegmentierung, Repräsentanz und Data Mining - Konzepte der digitalen Bodenkartierung. Mitteilgn. Dtsch. Bodenkundl. Gesellsch. 110(2): 537f.

Schmidt, K. Behrens, T. & Scholten T. (2005). Entwicklung eines Verfahrens zur Verschneidung von Attributinformationen bei nicht deckungsgleichen Bodeninformationen. IN: Tübinger geowissenschaftliche Arbeiten (TGA).

Behrens, T., **Schmidt, K.** Kipka, H. und Scholten, T. (2005). Prognose und Korrektur von Bodenkarten mit Techniken des Data. Minings. Mitteilgn. Dtsch. Bodenkundl. Gesellsch.

Schmidt, K. Behrens, T., Scholten, T. , Reinhardt, F. und W. Brandtner (2005). Räumliche Zuweisung und Extrapolation von Attributen der Mittelmaßstäbigen Bodenkarte (MMK) in die Bodengeologische Karte (BGK) Thüringens. Mitteilgn. Dtsch. Bodenkundl. Gesellsch. 106: 95f.